Ubiquitin-specific protease

Field of the invention

The present invention relates to ubiquitin-specific proteases (USPs), specifically, to novel members of the family of deubiquitinating enzymes (DUBs).

Background of the invention

Ubiquitin is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis. Conjugation to the small eukaryotic protein ubiquitin can functionally modify or target proteins for degradation by the proteasome. Like protein phosphorylation, protein ubiquitination is dynamic, involving enzymes that add ubiquitin (ubiquitin conjugating enzymes) and enzymes that remove ubiquitin. Removal of the ubiquitin modification, or deubiquitination, is performed by enzymes termed ubiquitin-specific proteases (USPs) or ubiquitin C-terminal hydrolases (UCHs) and is an important mechanism regulating this pathway. These enzymes can cleave either peptide bonds linking ubiquitin as part of a precursor fusion protein, releasing free ubiquitin moieties, or cleave bonds conjugating ubiquitin (post-translationally) to proteins.

Deubiquitinating enzymes are cysteine proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. There are two distinct families of deubiquitinating enzymes (DUBs). The first class consists of enzymes of about 25 Kd and is currently represented in human by UCHL1, UCHL3, Bap1 and UCH37. These proteins belong to family C12 in the classification of peptidase. The second family consist of large proteins (800 to 2000 residues) that share two regions of similarity, a region that contains a conserved cysteine which is probably implicated in the catalytic mechanism (cysteine box) and a region that contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism (histidine box). These proteins were first characterized in yeast and belong to family C19 in MEROPS and are represented by the USPs.

Deubiquitinating enzymes have multiple roles within the cell, including stabilization of some ubiquitin (Ub) conjugated substrates, degradation of other Ub-conjugated substrates and

recycling of the cell's free monomeric Ub pool. Some deubiquitinating enzymes remove Ub from cellular target proteins and thereby prevent proteasome mediated degradation (UBP2). Other deubiquitinating enzymes remove Ub from Ub-peptide degradation products produced by the proteasomes and thereby accelerate proteasome mediated degradation (Doa-4).

Recent efforts exploring sequencing of different genomes have increased the number of sequences displaying conserved features of the family. In *S. cerevisiae* for example, a family of 17 DUB enzymes can be identified in its completely sequenced genome. Several other proteins from higher eukaryotes that contain the conserved sequence motifs (cysteine and histidine boxes) have also been identified. However, most of these new sequences represent truncated members and as a consequence, real assignment as a deubiquitinating family member is difficult.

Summary of the Invention

In a first aspect, the invention provides an isolated DNA comprising a nucleotide sequence encoding a ubiquitin-specific protease selected from the group of:

(a) SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10

(b) a fragment of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10

(c) a derivative of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10

(d) a substantially homologous sequence of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10.

In one aspect of the invention, the DNA is selected from the group of SEQ ID No. 2, SEQ ID No. 6 or SEQ.ID No. 10.

In another aspect of the invention, the fragment is a fragment of SEQ ID No. 2, SEQ ID No. 6, or SEQ.ID No. 10 and retains ubiquitin-specific functional activity. The fragment may be at least 50, optionally at least 75 or at least 100 consecutive nucleotides.

In yet another aspect of the invention, a derivative of SEQ ID No. 2, SEQ ID No. 6, or SEQ.ID No. 10 is provided wherein deletions, additions or substitutions of amino acid residues within the amino acid sequence produce a functionally equivalent amino acid sequence which retains ubiquitin-specific functional activity.

In a further aspect of the invention, a nucleotide sequence is provided which is substantially homologous to a nucleotide sequence selected from the group consisting of SEQ. ID No. 2, SEQ ID No. 6, or SEQ.ID No. 10 and retains ubiquitin-specific functional activity. The percentage of homology between the substantially homologous sequence and the sequence SEQ. ID No. 2, SEQ ID No. 6, or SEQ.ID No. 10 desirably is at least 80%,

more desirably at least 85%, preferably at least 90%, more preferably at least 95%, still more preferably at least 99%.

In yet another aspect of the invention, a complementary nucleic acid sequence which hybridizes under high stringency conditions to an isolated DNA of any one of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10 or a fragment, derivative or substantially homologous sequence thereof is provided.

In a second aspect of the invention, an isolated polypeptide of an ubiquitin-specific protease is provided, comprising the amino acid sequence selected from the group of:

(a) SEQ.ID. No.1, SEQ.ID.No.5, SEQ.ID No. 9

(b) a fragment of SEQ.ID. No.1, SEQ.ID.No.5, SEQ.ID No. 9

(c) a derivative of SEQ.ID. No.1, SEQ.ID.No.5, SEQ.ID No. 9

(d) a substantially homologous sequence of SEQ.ID. No.1, SEQ.ID.No.5, SEQ.ID No. 9.

One preferred aspect of the invention provides an isolated polypeptide with an amino acid sequence as set forth in SEQ ID NO:1. Such a polypeptide, or fragments thereof, is found in the breast tissue of sufferers of breast cancer to a much greater extent than in the breast tissue of individuals without breast cancer. In accordance with this aspect of the invention there are provided novel polypeptides of human origin as well as biologically, diagnostically or therapeutically useful fragments, derivatives, and homologues of the foregoing.

Another preferred aspect of the invention provides an isolated polypeptide with an amino acid sequence as set forth in SEQ ID NO:5. Such a polypeptide, or fragments thereof, is found in the peripheral blood cells, especially lymphoid cells of sufferers of leukemia to a much greater extent than in the peripheral blood of individuals without leukemia. In accordance with this aspect of the invention there are provided novel polypeptides of human origin as well as biologically, diagnostically or therapeutically useful fragments, derivatives, and homologues of the foregoing.

Yet another preferred aspect of the invention provides an isolated polypeptide with an amino acid sequence as set forth in SEQ ID NO:9. Such a polypeptide, or fragments thereof, is found in the brain tissue, especially amygdale, spinal cord and olfactory bulb tissue of sufferers of brain disorders to a much greater extent than in the brain tissue of individuals without brain disorders. In accordance with this aspect of the invention there are provided novel polypeptides of human origin as well as biologically, diagnostically or therapeutically useful fragments, derivatives, and homologues of the foregoing.

A third aspect of the present invention encompasses a method for the diagnosis of in a human which requires measuring the amount of a polypeptide selected from the group of SEQ ID. No.1, SEQ ID. No.5 or SEQ ID No.9, a fragment, derivative or substantially homologous sequence thereof from a human, where the presence of an elevated amount of the polypeptide or fragments thereof, relative to the amount of the polypeptide or fragments thereof in normal tissue is diagnostic of the human's suffering from a disease.

In a one preferred aspect, a method for the diagnosis of breast cancer in a human comprising measuring the amount of a polypeptide according to SEQ ID. No.1, a fragment, derivative or substantially homologous sequence thereof from a human, in breast tissue, wherein the presence of an elevated amount of said polypeptide relative to the amount of said polypeptide in normal breast tissue is diagnostic of said human's suffering from breast cancer.

In another preferred aspect, a method for the diagnosis of leukemia in a human which comprises measuring the amount of a polypeptide according to SEQ ID. No.5, a fragment, derivative or substantially homologous sequence thereof from a human, in peripheral blood cells, especially lymphoid cells, wherein the presence of an elevated amount of said polypeptide relative to the amount of said polypeptide in normal peripheral blood cell, especially lymphoid cells is diagnostic of said human's suffering from leukemia.

In yet a further preferred aspect, a method for the diagnosis of brain disorders in a human which comprises measuring the amount of a polypeptide that comprises a polypeptide according to SEQ ID. No.9 a fragment, derivative or substantially homologous sequence thereof from a human, in the amgdala, spinal cord or olfactory bulb tissues, wherein the presence of an elevated amount of said polypeptide relative to the amount of said polypeptide in amgdala, spinal cord and olfactory bulb tissues is diagnostic of said human's suffering from a brain disorder.

Another aspect of the invention provides a process for producing the aforementioned polypeptides, polypeptide fragments, derivatives, homologues, fragments of the variants and derivatives, and homologs of the foregoing. In a preferred embodiment of this aspect of the invention there are provided methods for producing the aforementioned human polypeptides comprising culturing host cells having incorporated therein an expression vector containing an exogenously-derived ubiquitin-specific polynucleotides under conditions sufficient for expression of ubiquitin specific polypeptides in the host and then recovering the expressed polypeptide.

In accordance with another aspect of the invention there are provided products, compositions, processes and methods that utilize the aforementioned polypeptides and polynucleotides for, *inter alia*, research, biological, clinical and therapeutic purposes.

In certain additional preferred embodiments of this aspect of the invention there are provided an antibody or a fragment thereof which specifically binds to a polypeptide that comprises the amino acid sequence set forth in SEQ ID NO:1, SEQ ID. No.5 or SEQ ID. No.9. In certain particularly preferred embodiments in this regard, the antibodies are highly selective for human ubiquitin-specific polypeptides or portions of human ubiquitin-specific polypeptides.In a further aspect, an antibody or fragment thereof is provided that binds to a fragment of the amino acid sequence set forth in SEQ ID NO:1, SEQ ID. No.5 or SEQ ID. No.9. In a related aspect, a pharmaceutical composition comprising such an antibody is provided.

In another aspect, methods of treating a disease in a subject, where the disease is mediated by or associated with an increase in the presence of polypeptide of SEQ ID. No.1, SEQ ID. No.5 or SEQ ID. No.9 in breast tissue, peripheral blood cells, or brain tissue respectively, by the administration of an effective amount of an antibody that binds to a polypeptide with the amino acid sequence set out in SEQ ID NO:1, SEQ ID. No.5 or SEQ ID. No.9 or a fragment or portion thereof to the subject is provided. Also provided are methods for the diagnosis of a disease or condition associated with an increase in the presence of polypeptide in a subject, which comprises utilizing an antibody that binds to a polypeptide with the amino acid sequence set out in SEQ ID NO:1, SEQ ID. No.5 or SEQ ID. No.9, or a fragment or portion thereof in an immunoassay.

In yet another aspect, the invention provides cells which can be propagated in vitro, preferably mammalian, more preferably vertebrate cells, which are capable upon growth in culture of producing a polypeptide that comprises the amino acid sequence set forth in SEQ ID NO:1, SEQ ID. No.5 or SEQ ID. No.9 or fragments, derivatives or substantially homologous sequences thereof, where the cells optionally contain transcriptional control DNA sequences, other than human transcriptional control sequences, where the transcriptional control sequences control transcription of DNA encoding a polypeptide with the amino acid sequence.

In another aspect, the present invention provides a method for producing polypeptides which comprises culturing a host cell having incorporated therein an expression vector containing an exogenously-derived ubiquitin-specific polynucleotide of the invention under conditions sufficient for expression of such polypeptides in the host cell, thereby

causing the production of an expressed polypeptide, and recovering the expressed polypeptide.

In yet another aspect of the present invention there are provided assay methods and kits comprising the components necessary to detect above-normal expression of ubiquitin-specific polynucleotides of the invention or polypeptides or fragments thereof in body tissue samples derived from a patient, such kits comprising e.g., antibodies or oligonucleotide probes that hybridize with polynucleotides of the invention. In a preferred embodiment, such kits also comprise instructions detailing the procedures by which the kit components are to be used.

Another aspect is directed to pharmaceutical compositions comprising a nucleotide sequence of SEQ ID NO:1, SEQ ID. No.5 or SEQ ID. No.9, or a fragment, derivative or homologue thereof.

In another aspect, the invention is directed to methods for the identification of molecules that can bind to the ubiquitin-specific proteases of the invention and/or modulate the activity of ubiquitin or molecules that can bind to nucleic acid sequences that modulate the transcription or translation of ubiquitin. Such methods are disclosed in, e.g., U.S. Patent Nos. 5,541,070; 5,567,317; 5,593,853; 5,670,326; 5,679,582; 5,856,083; 5,858,657; 5,866,341; 5,876,946; 5,989,814; 6,010,861; 6,020,141; 6,030,779; and 6,043024, all of which are incorporated by reference herein in their entirety. Molecules identified by such methods also fall within the scope of the present invention.

In yet another aspect, the invention is directed to methods for the introduction of nucleic acids of the invention into one or more tissues of a subject in need of treatment with the result that one or more proteins encoded by the nucleic acids are expressed and or secreted by cells within the tissue.

Other aspects, features, advantages and aspects of the present invention will become apparent to those of skill from the following description. It should be understood, however, that the following description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only. Various changes and modifications within the spirit and scope of the disclosed invention will become readily apparent to those skilled in the art from reading the following description and from reading the other parts of the present disclosure.

Description of Tables

Table 1(a): USP_N01: Novel splice variant –amino acid sequence

Table 1(a) depicts SEQ ID. No.1 which is a novel splice form of a ubiquitin-specific protease. This novel splice form is characterized by 4 insertions at the following locations: Pos1- Pos11, Pos267 - Pos300, Pos361 - Pos384 and Pos1243 -Pos1437.

Table 1(b): USP_N01: Novel splice variant – nucleotide sequence

Table 1(b) depicts SEQ ID. No.2 which is the corresponding nucleotide sequence to Table 1(a).

Table 1(c): DERWENT reference amino acid sequence AAU82706

Table 1(c) depicts SEQ ID. No. 3 which is the reference amino acid sequence to which the novel splice form of Table 1(a) has been compared.

Table 1(d): reference nucleotide sequence

Table 1(d) depicts SEQ ID. No. 4 which is the reference nucleotide sequence corresponding to Table 1(c).

Table 2(a): USP_N07: Novel splice variant –amino acid sequence

Table 2(a) depicts SEQ ID. No.5 which is a novel splice form of a ubiquitin-specific protease. This novel splice form is characterized by 1 insertions at the following location: Pos14 – Pos81.

Table 2(b): USP_N07: Novel splice variant – nucleotide sequence

Table 2(b) depicts SEQ ID. No.6 which is the corresponding nucleotide sequence to Table 2(a).

Table 2(c): DERWENT reference amino acid sequence AAU82714

Table 2(c) depicts SEQ ID. No. 7 which is the reference amino acid sequence to which the novel splice form of Table 2(a) has been compared.

Table 2(d): reference nucleotide sequence

Table 2(d) depicts SEQ ID. No. 8 which is the reference nucleotide sequence corresponding to Table 2(c).

Table 3(a): USP_N11: Novel splice variant –amino acid sequence

Table 3(a) depicts SEQ ID. No.9 which is a novel splice form of a ubiquitin-specific protease. This novel splice form is characterized by 1 insertions at the following location: Pos12 – Pos48.

Table 3(b): USP_N11: Novel splice variant – nucleotide sequence

Table 3(b) depicts SEQ ID. No.10 which is the corresponding nucleotide sequence to Table 3(a).

Table 3(c): DERWENT reference amino acid sequence AAU82713

Table 3(c) depicts SEQ ID. No.11 which is the reference amino acid sequence to which the novel splice form of Table 3(a) has been compared.

Table 3(d): reference nucleotide sequence

Table 3(d) depicts SEQ ID. No.12 which is the reference nucleotide sequence corresponding to Table 3(c).

Detailed Description of the Invention

All patent applications, patents and literature references cited herein are hereby incorporated by reference in their entirety.

In practicing the present invention, many conventional techniques in molecular biology, microbiology, and recombinant DNA are used. These techniques are well known and are explained in, for example, Current Protocols in Molecular Biology, Volumes I, II, and III, 1997 (F. M. Ausubel ed.); Sambrook et al., 1989, Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; DNA Cloning: A Practical Approach, Volumes I and II, 1985 (D. N. Glover ed.); Oligonucleotide Synthesis, 1984 (M. L. Gait ed.); Nucleic Acid Hybridization, 1985, (Hames and Higgins); Transcription and Translation, 1984 (Hames and Higgins eds.); Animal Cell Culture, 1986 (R. I. Freshney ed.); Immobilized Cells and Enzymes, 1986 (IRL Press); Perbal, 1984, A Practical Guide to Molecular Cloning; the series, Methods in Enzymology (Academic Press, Inc.); Gene Transfer Vectors for Mammalian Cells, 1987 (J. H. Miller and M. P. Calos eds., Cold Spring Harbor Laboratory); and Methods in Enzymology Vol. 154 and Vol. 155 (Wu and Grossman, and Wu, eds., respectively).

The invention relates to the identification of novel splice variants of human ubiquitin-specific proteases known as DUBs or deubiquitinating enzymes. The novel splice variants of the invention are provided in Tables 1(a), 2(a) and 3(a) which correspond to the amino acid sequences SEQ ID. No.1, SEQ ID No.5 and SEQ ID. No. 9 respectively. The corresponding nucleotide sequences are provided in Tables 1(b), 2(b) and 3(b) which are SEQ ID No. 2, SEQ ID No. 6 and SEQ ID No. 10 respectively. These sequences are splice variants of the DERWENT reference sequences AAU82706, AAU82714 and AAU82713 respectively.

The invention encompasses nucleic acid sequences and amino acid sequences which are substantially homologous to the sequences provided in Tables 1,2 and 3.

However, it is understood that any amino acid sequences disclosed prior to this invention are excluded. The term "substantially homologous", when used herein with respect to a sequence means that a sequence when compared to its corresponding reference sequence, has substantially the same structure and function. When a position in the reference sequence is occupied by the same amino acid or nucleotide the molecules are homologous at that position (i.e. there is identity at that position). In the case of nucleic acid sequence comparison there is also homology at a certain position where the codon triplet including the nucleotide encodes the same amino acid in both molecules being compared due to degeneracy of the genetic code.

The percentage of homology between the substantially homologous sequence and the reference sequence desirably is at least 80%, more desirably at least 85%, preferably at least 90%, more preferably at least 95%, still more preferably at least 99%.

Sequence comparisons are carried out using a Smith-Waterman sequence alignment algorithm (see e.g. Waterman, M.S. Introduction to Computational Biology: Maps, sequences and genomes. Chapman & Hall. London: 1995. ISBN 0-412-99391-0, or at http://www-to.usc.edu/software/seqaln/index.html).

Also comprised within the nucleic acid sequences of the present invention are sequences which hybridize to the nucleic acid sequences of the present invention under stringent hybridization conditions. Stringent hybridization conditions are defined as a positive hybridization signal observed after washing in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 2X SSC, 0.1% SDS at 50°C, more desirably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 1X SSC, 0.1% SDS at 50°C, still more desirably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 0.5X SSC, 0.1% SDS at 50°C, preferably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 0.1X SSC, 0.1% SDS at 50°C, more preferably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 0.1X SSC, 0.1% SDS at 65°C.

The nucleic acid sequences of the invention may incorporate the Open Reading Frame (ORF) and may also incorporate the 5' Untranslated Region (UTR) or a portion thereof. The invention may also include any promoter, enhancer, regulatory, terminator and localization elements and use of these elements in conjunction with heterlogous genes.

The definition of homologous sequences provided above embraces fragments of the novel splice variants of the nucleic acid or amino acid sequences of the present invention.

A "fragment" means any peptide molecule having at least 5, 10, 15 or optionally at least 25,35, or 45 contiguous amino acids of the novel splice variant. A fragment of a nucleic acid sequence comprises at least 50, optionally at least 75 or at least 100 consecutive nucleotides.

The fragments to which the invention pertains, however, are not to be construed as encompassing fragments that may have been disclosed prior to the present invention. Fragments can retain one or more of the biological activities of the protein, for example the ability to bind to ubiquitin or hydrolyze peptide bonds, as well as fragments that can be used as an immunogen to generate ubiquitin protease antibodies. Biologically active fragments can comprise a domain or motif, e.g., catalytic site, UBP or UCH signature, membrane-associated regions and sites for glycosylation, cAMP and cGMP-dependent protein kinase phosphorylation, protein kinase C phosphorylation, casein kinase II phosphorylation, tyrosine kinase phosphorylation, N-myristoylation, and amidation.

Further possible fragments include the catalytic site or domain including the cysteine or histidine ubiquitin recognition sites, ubiquitin binding sites, sites important for subunit interaction, and sites important for carrying out the other functions of the protease. Such domains or motifs can be identified by means of routine computerized homology searching procedures. Fragments, for example, can extend in one or both directions from the functional site to encompass 5, 10, 15, 20, 30, 40, 50, or up to 100 amino acids.

Further, fragments can include sub-fragments of the specific domains mentioned above, which sub-fragments retain the function of the domain from which they are derived. These regions can be identified by well-known methods involving computerized homology analysis. The invention also provides fragments with immunogenic properties. These contain an epitope-bearing portion of the ubiquitin protease and variants. These epitope-bearing peptides are useful to raise antibodies that bind specifically to a ubiquitin protease polypeptide or region or fragment. These peptides can contain at least 10, 12, at least 14, or between at least about 15 to about 30 amino acids. Non-limiting examples of antigenic polypeptides that can be used to generate antibodies include but are not limited to peptides derived from an extracellular site with regions having a high antigenicity index (see FIG. 3 US 6,451,994). However, intracellularly-made antibodies ("intrabodies") are also encompassed, which would recognize intracellular peptide regions. The epitope-bearing ubiquitin protease polypeptides may be produced by any conventional means (Houghten, R.

A. (1985) Proc. Natl. Acad. Sci. USA 82:5131-5135). Simultaneous multiple peptide synthesis is described in U.S. Pat. No. 4,631,211.

Fragments can be discrete (not fused to other amino acids or polypeptides) or can be within a larger polypeptide. Further, several fragments can be comprised within a single larger polypeptide. In one embodiment a fragment designed for expression in a host can have heterologous pre- and pro-polypeptide regions fused to the amino terminus of the ubiquitin protease fragment and an additional region fused to the carboxyl terminus of the fragment.

Fragments may be used as a hybridization probe for a cDNA library to isolate the full length gene and to isolate other genes which have a high sequence similarity or similar biological activity. Probes of this type preferably have at least about 30 bases and may contain, for example, from about 30 to about 50 bases, about 50 to about 100 bases, about 100 to about 200 bases, or more than 200 bases. The probe may also be used to identify a cDNA clone corresponding to a full length transcript and a genomic clone or clones that contain the full-length gene including regulatory and promoter regions, exons, and introns. An example of a screen comprises isolating the coding region of the gene by using the known DNA sequence to synthesize an oligonucleotide probe. Labeled oligonucleotides having a sequence complementary to that of the gene of the present invention are used to screen a library of human cDNA, genomic DNA or mRNA to determine which members of the library the probe hybridizes to.

The present invention also encompasses "derivatives" of the amino acid sequences. A "derivative" is a sequence related to the amino acid sequence either on the amino acid level (e.g. a homologous sequence wherein certain naturally-occurring amino acids are replaced with synthetic amino acid substitutes or at the three dimensional level (e.g. wherein molecules have approximately the same shape and conformation as the amino acid sequence. Thus derivatives include mutants, mimetics, mimotopes, analogues, monomeric forms and functional equivalents.

Deletions, additions or substitutions of amino acid residues within the amino acid sequence which result in a silent change produce a functionally equivalent differentially expressed gene product. Amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues involved. For example, nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine; polar neutral

amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine; positively charged (basic) amino acids include arginine, lysine, and histidine; and negatively charged (acidic) amino acids include aspartic acid and glutamic acid.

Derivatives may include those in which a substituted amino acid residue is not one encoded by the genetic code, in which a substituent group is included, in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or in which the additional amino acids are fused to the mature polypeptide, such as a leader or secretory sequence or a sequence for purification of the mature polypeptide or a pro-protein sequence.

Known modifications include, but are not limited to, acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphatidylinositol, cross-linking, cyclization, disulfide bond formation, demethylation, formation of covalent crosslinks, formation of cystine, formation of pyroglutamate, formylation, gamma carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination.

Such modifications are well-known to those of skill in the art and have been described in great detail in the scientific literature. Several particularly common modifications, glycosylation, lipid attachment, sulfation, gamma-carboxylation of glutamic acid residues, hydroxylation and ADP-ribosylation, for instance, are described in most basic texts, such as Proteins--Structure and Molecular Properties, 2nd ed., T. E. Creighton, W.H. Freeman and Company, New York (1993). Many detailed reviews are available on this subject, such as by Wold, F., Posttranslational Covalent Modification of Proteins, B. C. Johnson, Ed., Academic Press, New York 1-12 (1983); Seifter et al. (1990) Meth. Enzymol. 182: 626-646) and Rattan et al. (1992) Ann. N.Y. Acad. Sci. 663:48-62).

As is also well known, polypeptides are not always entirely linear. For instance, polypeptides may be branched as a result of ubiquitination, and they may be circular, with or without branching, generally as a result of post-translation events, including natural processing events and events brought about by human manipulation which do not occur naturally. Circular, branched and branched circular polypeptides may be synthesized by non-translational natural processes and by synthetic methods.

Modifications can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. Blockage of the amino or carboxyl group in a polypeptide, or both, by a covalent modification, is common in naturally-occurring and synthetic polypeptides. For instance, the aminoterminal residue of polypeptides made in E. coli, prior to proteolytic processing, almost invariably will be N-formylmethionine.

The modifications can be a function of how the protein is made. For recombinant polypeptides, for example, the modifications will be determined by the host cell posttranslational modification capacity and the modification signals in the polypeptide amino acid sequence. Accordingly, when glycosylation is desired, a polypeptide should be expressed in a glycosylating host, generally a eukaryotic cell. Insect cells often carry out the same posttranslational glycosylations as mammalian cells and, for this reason, insect cell expression systems have been developed to efficiently express mammalian proteins having native patterns of glycosylation. Similar considerations apply to other modifications.The same type of modification may be present in the same or varying degree at several sites in a given polypeptide. Also, a given polypeptide may contain more than one type of modification.

"Functionally equivalent," as utilized herein, may refer to a protein or polypeptide capable of exhibiting a substantially similar *in vivo* or *in vitro* activity as the endogenous differentially expressed gene products encoded by the differentially expressed gene sequences described above. "Functionally equivalent" may also refer to proteins or polypeptides capable of interacting with other cellular or extracellular molecules in a manner substantially similar to the way in which the corresponding portion of the endogenous differentially expressed gene product would. For example, a "functionally equivalent" peptide would be able, in an immunoassay, to diminish the binding of an antibody to the corresponding peptide (i.e., the peptide the amino acid sequence of which was modified to achieve the "functionally equivalent" peptide) of the endogenous protein, or to the endogenous protein itself, where the antibody was raised against the corresponding peptide of the endogenous protein. An equimolar concentration of the functionally equivalent peptide will diminish the aforesaid binding of the corresponding peptide by at least about 5%, preferably between about 5% and 10%, more preferably between about 10% and 25%, even more preferably between about 25% and 50%, and most preferably between about 40% and 50%.

For example, functionally equivalent peptides can be fully functional or can lack function in one or more activities. Thus, in the present invention, variations can affect the function, for example, of ubiquitin binding, ubiquitin recognition, interaction with ubiquitinated substrate protein, such as binding or proteolysis, subunit interaction, particularly within the proteasome, activation or binding by ATP, developmental expression, temporal expression, tissue-specific expression, interacting with cellular components, such as transcriptional regulatory factors, and particularly trans-acting transcriptional regulatory factors, proteolytic cleavage of peptide bonds in polyubiquitin and peptide bonds between ubiquitin or polyubiquitin and substrate protein, and proteolytic cleavage of peptide bonds between ubiquitin or polyubiquitin and a peptide or amino acid.

Fully functional variants typically contain only conservative variation or variation in non-critical residues or in non-critical regions. Functional variants can also contain substitution of similar amino acids, which results in no change or an insignificant change in function. Alternatively, such substitutions may positively or negatively affect function to some degree.

Non-functional variants typically contain one or more non-conservative amino acid substitutions, deletions, insertions, inversions, or truncation or a substitution, insertion, inversion, or deletion in a critical residue or critical region. As indicated, variants can be naturally-occurring or can be made by recombinant means or chemical synthesis to provide useful and novel characteristics for the ubiquitin protease polypeptide. This includes preventing immunogenicity from pharmaceutical formulations by preventing protein aggregation. Useful variations further include alteration of catalytic activity. For example, one embodiment involves a variation at the binding site that results in binding but not hydrolysis, or slower hydrolysis, of the peptide bond. A further useful variation results in an increased rate of hydrolysis of the peptide bond. A further useful variation at the same site can result in higher or lower affinity for substrate.

Useful variations also include changes that provide for affinity for a different ubiquitinated substrate protein than that normally recognized. Other useful variations involving altered recognition affect recognition of the type of substrate normally recognized. For example, one variation could result in recognition of ubiquitinated intact substrate but not of substrate remnants, such as ubiquitinated amino acid or peptide that are proteolysis products that result from the hydrolysis of the intact ubiquitinated substrate. Alternatively, the protease could be varied so that one or more of the remnant products is recognized but not the intact protein substrate. Another variation would affect the ability of the protease to rescue a ubiquitinated protein. Thus, protein substrates that are normally rescued from

proteolysis would be subject to degradation. Further useful variations affect the ability of the protease to be induced by activators, such as cytokines, including but not limited to, those disclosed herein. Another useful variation would affect the recognition of ubiquitin substrate so that the enzyme could not recognize one or more of a linear polyubiquitin, branched chain polyubiquitin, linear polyubiquitinated substrate, or branched chain polyubiquitin substrate. Specific variations include truncation in which, for example, a HIS domain is deleted, the variation resulting in decrease or loss of deubiquitination activity. Another useful variation includes one that prevents activation by ATP.

Another useful variation provides a fusion protein in which one or more domains or subregions are operationally fused to one or more domains or subregions from another UBP or from a UCH. Specifically, a domain or subregion can be introduced that provides a rescue function to an enzyme not normally having this function or for recognition of a specific substrate wherein recognition is not available to the original enzyme. Other variations include those that affect ubiquitin recognition or recognition of a ubiquitinated substrate protein. Further variations could affect specific subunit interaction, particularly in the proteasome. Other variations would affect developmental, temporal, or tissue-specific expression. Other variations would affect the interaction with cellular components, such as transcriptional regulatory factors. Amino acids that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine- scanning mutagenesis (Cunningham et al. (1985) Science 244:1081- 1085). The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity, such as peptide hydrolysis in vitro or ubiquitin- dependent in vitro activity, such as proliferative activity, receptor- mediated signal transduction, and other cellular processes including, but not limited, those disclosed herein that are a function of the ubiquitin system. Sites that are critical for binding or recognition can also be determined by structural analysis such as crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith et al. (1992) J. Mol. Biol. 224:899-904; de Vos et al. (1992) Science 255:306-312).

The assays for deubiquitinating enzyme activity are well known in the art and can be found, for example, in Zhu et al. (1997) Journal of Biological Chemistry 272:51-57, Mitch et al. (1999) American Journal of Physiology 276:C1132-C1138, Liu et al. (1999) Molecular and Cell Biology 19:3029-3038, and such as those cited in various reviews, for example, Ciechanover et al. (1994) The FASEB Journal 8:182-192, Chiechanover (1994) Biol. Chem. Hoppe-Seyler 375:565- 581, Hershko et al. (1998) Annual Review of Biochemistry 67:425-

479, Swartz (1999) Annual Review of Medicine 50:57-74, Ciechanover (1998) EMBO Journal 17:7151-7160, and D'Andrea et al. (1998) Critical Reviews in Biochemistry and Molecular Biology 33:337-352. These assays include, but are not limited to, the disappearance of substrate, including decrease in the amount of polyubiquitin or ubiquitinated substrate protein or protein remnant, appearance of intermediate and end products, such as appearance of free ubiquitin monomers, general protein turnover, specific protein turnover, ubiquitin binding, binding to ubiquitinated substrate protein, subunit interaction, interaction with ATP, interaction with cellular components such as trans- acting regulatory factors, stabilization of specific proteins, and the like.

A "host cell," as used herein, refers to a prokaryotic or eukaryotic cell that contains heterologous DNA that has been introduced into the cell by any means, e.g., electroporation, calcium phosphate precipitation, microinjection, transformation, viral infection, and the like.

"Heterologous" as used herein means "of different natural origin" or represent a non-natural state. For example, if a host cell is transformed with a DNA or gene derived from another organism, particularly from another species, that gene is heterologous with respect to that host cell and also with respect to descendants of the host cell which carry that gene. Similarly, heterologous refers to a nucleotide sequence derived from and inserted into the same natural, original cell type, but which is present in a non-natural state, e.g. a different copy number, or under the control of different regulatory elements.

A "vector" is a nucleic acid molecule into which heterologous nucleic acid may be inserted which can then be introduced into an appropriate host cell. Vectors preferably have one or more origin of replication, and one or more site into which the recombinant DNA can be inserted. Vectors often have convenient means by which cells with vectors can be selected from those without, e.g., they encode drug resistance genes. Common vectors include plasmids, viral genomes, and (primarily in yeast and bacteria) "artificial chromosomes."

"Plasmids" generally are designated herein by a lower case p preceded and/or followed by capital letters and/or numbers, in accordance with standard naming conventions that are familiar to those of skill in the art. Starting plasmids disclosed herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids by routine application of well known, published procedures. Many plasmids and other cloning and expression vectors that can be used in accordance with the present invention are well known and readily available to those of skill in the art. Moreover, those of skill readily may construct any number of other plasmids suitable for use in the

invention. The properties, construction and use of such plasmids, as well as other vectors, in the present invention will be readily apparent to those of skill from the present disclosure.

The term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated, even if subsequently reintroduced into the natural system. Such polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

As used herein, the term "transcriptional control sequence" refers to DNA sequences, such as initiator sequences, enhancer sequences, and promoter sequences, which induce, repress, or otherwise control the transcription of protein encoding nucleic acid sequences to which they are operably linked. "Human transcriptional control sequences" are any of those transcriptional control sequences naturally occurring in the human genome whereas "non-human transcriptional control sequences" are any transcriptional control sequences not found in the human genome.

A variety of host-expression vector systems may be utilized to express the gene coding sequences of the invention. Such host-expression systems represent vehicles by which the coding sequences of interest may be produced and subsequently purified, but also represent cells which may, when transformed or transfected with the appropriate nucleotide coding sequences, exhibit the differentially expressed gene protein of the invention *in situ*. These include but are not limited to microorganisms such as bacteria (e.g., E. coli, B. subtilis) transformed with recombinant bacteriophage DNA, plasmid DNA or cosmid DNA expression vectors containing differentially expressed gene protein coding sequences; yeast (e.g. Saccharomyces, Pichia) transformed with recombinant yeast expression vectors containing the differentially expressed gene protein coding sequences; insect cell systems infected with recombinant virus expression vectors (e.g., baculovirus) containing the differentially expressed gene protein coding sequences; plant cell systems infected with recombinant virus expression vectors (e.g., cauliflower mosaic virus, CaMV; tobacco mosaic virus, TMV) or transformed with recombinant plasmid transformation vectors (e.g., Ti plasmid) containing differentially expressed gene protein coding sequences; or mammalian cell systems (e.g. COS, CHO, BHK, 293, 3T3) harboring recombinant expression constructs containing promoters derived from the genome of mammalian cells (e.g., metallothioneine

promoter) or from mammalian viruses (e.g., the adenovirus late promoter; the vaccinia virus 7.5K promoter).

In bacterial systems, a number of expression vectors may be advantageously selected depending upon the use intended for the differentially expressed gene protein being expressed. For example, when a large quantity of such a protein is to be produced, for the generation of antibodies or to screen peptide libraries, for example, vectors which direct the expression of high levels of fusion protein products that are readily purified may be desirable. Such vectors include, but are not limited, to the *E. coli* expression vector pUR278 (Ruther et al., 1983, EMBO J. 2:1791), in which the differentially expressed gene protein coding sequence may be ligated individually into the vector in frame with the lac Z coding region so that a fusion protein is produced; pIN vectors (Inouye & Inouye, 1985, Nucleic Acids Res. 13:3101-3109; Van Heeke & Schuster, 1989, J. Biol. Chem. 264:5503-5509); and the like. PGEX vectors may also be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST). In general, such fusion proteins are soluble and can easily be purified from lysed cells by adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione. The PGEX vectors are designed to include thrombin or factor Xa protease cleavage sites so that the cloned target gene protein can be released from the GST moiety.

Promoter regions can be selected from any desired gene using vectors that contain a reporter transcription unit lacking a promoter region, such as a chloramphenicol acetyl transferase ("cat") transcription unit, downstream of restriction site or sites for introducing a candidate promoter fragment; i.e., a fragment that may contain a promoter. As is well known, introduction into the vector of a promoter-containing fragment at the restriction site upstream of the cat gene engenders production of CAT activity, which can be detected by standard CAT assays. Vectors suitable to this end are well known and readily available. Two such vectors are pKK232-8 and pCM7. Thus, promoters for expression of polynucleotides of the present invention include not only well known and readily available promoters, but also promoters that readily may be obtained by the foregoing technique, using a reporter gene.

Among known bacterial promoters suitable for expression of polynucleotides and polypeptides in accordance with the present invention are the E. coli lacI and lacZ promoters, the T3 and T7 promoters, the T5 tac promoter, the lambda PR, PL promoters and the trp promoter. Among known eukaryotic promoters suitable in this regard are the CMV immediate early promoter, the HSV thymidine kinase promoter, the early and late SV40

promoters, the promoters of retroviral LTRs, such as those of the Rous sarcoma virus ("RSV"), and metallothionein promoters, such as the mouse metallothionein-I promoter.

In an insect system, *Autographa californica* nuclear polyhedrosis virus (AcNPV) is one of several insect systems that can be used as a vector to express foreign genes. The virus grows in Spodoptera frugiperda cells. The differentially expressed gene coding sequence may be cloned individually into non-essential regions (for example the polyhedrin gene) of the virus and placed under control of an AcNPV promoter (for example the polyhedrin promoter). Successful insertion of differentially expressed gene coding sequence will result in inactivation of the polyhedrin gene and production of non-occluded recombinant virus (i.e., virus lacking the proteinaceous coat coded for by the polyhedrin gene). These recombinant viruses are then used to infect Spodoptera frugiperda cells in which the inserted gene is expressed. (E.g., see Smith et al., 1983, J. Virol. 46: 584; Smith, U.S. Pat. No. 4,215,051).

In mammalian host cells, a number of viral-based expression systems may be utilized. In cases where an adenovirus is used as an expression vector, the differentially expressed gene coding sequence of interest may be ligated to an adenovirus transcription/translation control complex, e.g., the late promoter and tripartite leader sequence. This chimeric gene may then be inserted in the adenovirus genome by in vitro or in vivo recombination. Insertion in a non-essential region of the viral genome (e.g., region E1 or E3) will result in a recombinant virus that is viable and capable of expressing differentially expressed gene protein in infected hosts. (E.g., See Logan & Shenk, 1984, Proc. Natl. Acad. Sci. USA 81:3655-3659). Specific initiation signals may also be required for efficient translation of inserted differentially expressed gene coding sequences. These signals include the ATG initiation codon and adjacent sequences. In cases where an entire differentially expressed gene, including its own initiation codon and adjacent sequences, is inserted into the appropriate expression vector, no additional translational control signals may be needed. However, in cases where only a portion of the differentially expressed gene coding sequence is inserted, exogenous translational control signals, including, perhaps, the ATG initiation codon, must be provided. Furthermore, the initiation codon must be in phase with the reading frame of the desired coding sequence to ensure translation of the entire insert. These exogenous translational control signals and initiation codons can be of a variety of origins, both natural and synthetic. The efficiency of expression may be enhanced by the inclusion of appropriate transcription enhancer elements, transcription terminators, etc. (see Bittner et al., 1987, Methods in Enzymol. 153:516-544).

Selection of appropriate vectors and promoters for expression in a host cell is a well known procedure and the requisite techniques for expression vector construction, introduction of the vector into the host and expression in the host per se are routine skills in the art.

Generally, recombinant expression vectors will include origins of replication, a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence, and a selectable marker to permit isolation of vector containing cells after exposure to the vector.

In addition, a host cell strain may be chosen which modulates the expression of the inserted sequences, or modifies and processes the gene product in the specific fashion desired. Such modifications (e.g., glycosylation) and processing (e.g., cleavage) of protein products may be important for the function of the protein. Different host cells have characteristic and specific mechanisms for the post-translational processing and modification of proteins. Appropriate cell lines or host systems can be chosen to ensure the correct modification and processing of the foreign protein expressed. To this end, eukaryotic host cells which possess the cellular machinery for proper processing of the primary transcript, glycosylation, and phosphorylation of the gene product may be used. Such mammalian host cells include but are not limited to CHO, VERO, BHK, HeLa, COS, MDCK, 293, 3T3, WI38, etc.

For long-term, high-yield production of recombinant proteins, stable expression is preferred. For example, cell lines which stably express the differentially expressed gene protein may be engineered. Rather than using expression vectors which contain viral origins of replication, host cells can be transformed with DNA controlled by appropriate expression control elements (e.g., promoter, enhancer, sequences, transcription terminators, polyadenylation sites, etc.), and a selectable marker. Following the introduction of the foreign DNA, engineered cells may be allowed to grow for 1-2 days in an enriched media, and then are switched to a selective media. The selectable marker in the recombinant plasmid confers resistance to the selection and allows cells to stably integrate the plasmid into their chromosomes and grow to form foci which in turn can be cloned and expanded into cell lines. This method may advantageously be used to engineer cell lines which express the differentially expressed gene protein. Such engineered cell lines may be particularly useful in screening and evaluation of compounds that affect the endogenous activity of the differentially expressed gene protein.

A number of selection systems may be used, including but not limited to the herpes simplex virus thymidine kinase (Wigler, et al., 1977, Cell 11:223), hypoxanthine-guanine phosphoribosyltransferase (Szybalska & Szybalski, 1962, Proc. Natl. Acad. Sci. USA 48:2026), and adenine phosphoribosyltransferase (Lowy, et al., 1980, Cell 22:817) genes can be employed in tk^-, $hgprt^-$ or $aprt^-$ cells, respectively. Also, antimetabolite resistance can be used as the basis of selection for dhfr, which confers resistance to methotrexate (Wigler, et al., 1980, Natl. Acad. Sci. USA 77:3567; O'Hare, et al., 1981, Proc. Natl. Acad. Sci. USA 78:1527); gpt, which confers resistance to mycophenolic acid (Mulligan & Berg, 1981, Proc. Natl. Acad. Sci. USA 78:2072); neo, which confers resistance to the aminoglycoside G-418 (Colberre-Garapin, et al., 1981, J. Mol. Biol. 150:1); and hygro, which confers resistance to hygromycin (Santerre, et al., 1984, Gene 30:147) genes.

An alternative fusion protein system allows for the ready purification of non-denatured fusion proteins expressed in human cell lines (Janknecht, et al., 1991, Proc. Natl. Acad. Sci. USA 88: 8972-8976). In this system, the gene of interest is subcloned into a vaccinia recombination plasmid such that the gene's open reading frame is translationally fused to an amino-terminal tag consisting of six histidine residues. Extracts from cells infected with recombinant vaccinia virus are loaded onto Ni^{2+} nitriloacetic acid-agarose columns and histidine-tagged proteins are selectively eluted with imidazole-containing buffers.

When used as a component in assay systems such as those described below, the differentially expressed gene protein may be labeled, either directly or indirectly, to facilitate detection of a complex formed between the differentially expressed gene protein and a test substance. Any of a variety of suitable labeling systems may be used including but not limited to radioisotopes such as ^{125}I; enzyme labeling systems that generate a detectable calorimetric signal or light when exposed to substrate; and fluorescent labels.

Where recombinant DNA technology is used to produce the differentially expressed gene protein for such assay systems, it may be advantageous to engineer fusion proteins that can facilitate labeling, immobilization and/or detection.

Indirect labeling involves the use of a protein, such as a labeled antibody, which specifically binds to either a differentially expressed gene product. Such antibodies include but are not limited to polyclonal, monoclonal, chimeric, single chain, Fab fragments and fragments produced by an Fab expression library.

In another aspect, ubiquiting-specific protease polypeptides of the invention are useful in biological assays related to ubiquitin protease function. Such assays involve any of the known functions or activities or properties useful for diagnosis and treatment of ubiquitin-

or ubiquitin protease-related conditions. Potential assays have been disclosed herein and generically include disappearance of substrate, appearance of end product, and general or specific protein turnover.

The ubiquitin-specific protease polypeptides are also useful in drug screening assays, in cell-based or cell-free systems. Cell-based systems can be native, i.e., cells that normally express the ubiquitin protease, as a biopsy or expanded in cell culture. In one embodiment, however, cell- based assays involve recombinant host cells expressing the ubiquitin protease. Determining the ability of the test compound to interact with the ubiquitin protease can also comprise determining the ability of the test compound to preferentially bind to the polypeptide as compared to the ability of a known binding molecule (e.g., ubiquitin) to bind to the polypeptide. The polypeptides can be used to identify compounds that modulate ubiquitin protease activity. Such compounds, for example, can increase or decrease affinity for polyubiquitin, either linear or branched chain, ubiquitinated protein substrate, or ubiquitinated protein substrate remnants. Such compounds could also, for example, increase or decrease the rate of binding to these components. Such compounds could also compete with these components for binding to the ubiquitin protease or displace these components bound to the ubiquitin protease. Such compounds could also affect interaction with other components, such as ATP, other subunits, for example, in the 19S complex, and transcriptional regulatory factors. It is understood, therefore, that such compounds can be identified not only by means of ubiquitin, but by means of any of the components that functionally interact with the disclosed protease. This includes, but is not limited to, any of those components disclosed herein.

Ubiquitin-specific proteases, derivatives and fragments can be used in high-throughput screens to assay candidate compounds for the ability to bind to the ubiquitin protease. These compounds can be further screened against a functional ubiquitin protease to determine the effect of the compound on the ubiquitin protease activity. Compounds can be identified that activate (agonist) or inactivate (antagonist) the ubiquitin protease to a desired degree. Modulatory methods can be performed in vitro (e.g., by culturing the cell with the agent) or, alternatively, in vivo (e.g., by administering the agent to a subject.

The ubiquitin-specific protease polypeptides of the present inventioncan be used to screen a compound for the ability to stimulate or inhibit interaction between the ubiquitin protease protein and a target molecule that normally interacts with the ubiquitin protease protein. The target can be ubiquitin, ubiquitinated substrate, or polyubiquitin or another component of the pathway with which the ubiquitin protease protein normally interacts (for

example, ATP) . The assay includes the steps of combining the ubiquitin protease protein with a candidate compound under conditions that allow the ubiquitin protease protein or fragment to interact with the target molecule, and to detect the formation of a complex between the ubiquitin protease protein and the target or to detect the biochemical consequence of the interaction with the ubiquitin protease and the target. Any of the associated effects of protease function can be assayed. This includes the production of hydrolysis products, such as free terminal peptide substrate, free terminal amino acid from the hydrolyzed substrate, free ubiquitin, lower molecular weight species of hydrolyzed polyubiquitin, released intact substrate protein resulting from rescue from proteolysis, free polyubiquitin formed from hydrolysis of the polyubiquitin from intact substrate, and substrate remnants, such as amino acids and peptides produced from proteolysis of the substrate protein, and biological endpoints of the pathway.

Determining the ability of the ubiquitin protease to bind to a target molecule can also be accomplished using a technology such as real-time Bimolecular Interaction Analysis (BIA). Sjolander et al. (1991) Anal. Chem. 63:2338-2345 and Szabo et al. (1995) Curr. Opin. Struct. Biol. 5:699-705. As used herein, "BIA" is a technology for studying biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore®). Changes in the optical phenomenon surface plasmon resonance (SPR) can be used as an indication of real-time reactions between biological molecules. The test compounds of the present invention can be obtained using any of the numerous approaches in combinatorial library methods known in the art, including: biological libraries; spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; the 'one-bead one-compound' library method; and synthetic library methods using affinity chromatography selection. The biological library approach is limited to polypeptide libraries, while the other four approaches are applicable to polypeptide, non-peptide oligomer or small molecule libraries of compounds (Lam, K. S. (1997) Anticancer Drug Des. 12:145).

Examples of methods for the synthesis of molecular libraries can be found in the art, for example in DeWitt et al. (1993) Proc. Natl. Acad. Sci. USA 90:6909; Erb et al. (1994) Proc. Natl. Acad. Sci. USA 91:11422; Zuckermann et al. (1994). J. Med. Chem. 37:2678; Cho et al. (1993) Science 261:1303; Carell et al. (1994) Angew. Chem. Int. Ed. Engl. 33:2059; Carell et al. (1994) Angew. Chem. Int. Ed. Engl. 33:2061; and in Gallop et al. (1994) J. Med. Chem. 37:1233. Libraries of compounds may be presented in solution (e.g., Houghten (1992) Biotechniques 13:412-421), or on beads (Lam (1991) Nature 354:82-84), chips (Fodor (1993) Nature 364:555-556), bacteria (Ladner U.S. Pat. No. 5,223,409), spores

(Ladner U.S. Pat. No. '409), plasmids (Cull et al. (1992) Proc. Natl. Acad. Sci. USA 89:1865-1869) or on phage (Scott and Smith (1990) Science 249:386- 390); (Devlin (1990) Science 249:404-406); (Cwirla et al. (1990) Proc. Natl. Acad. Sci. 97:6378-6382); (Felici (1991) J. Mol. Biol. 222:301-310); (Ladner supra). Candidate compounds include, for example, 1) peptides such as soluble peptides, including Ig-tailed fusion peptides and members of random peptide libraries (see, e.g., Lam et al. (1991) Nature 354:82-84; Houghten et al. (1991) Nature 354:84-86) and combinatorial chemistry-derived molecular libraries made of D- and/or L-configuration amino acids; 2) phosphopeptides (e.g., members of random and partially degenerate, directed phosphopeptide libraries, see, e.g., Songyang et al. (1993) Cell 72:767-778); 3) antibodies (e.g., polyclonal, monoclonal, humanized, anti-idiotypic, chimeric, and single chain antibodies as well as Fab, F(ab') 2 , Fab expression library fragments, and epitope-binding fragments of antibodies); and 4) small organic and inorganic molecules (e.g., molecules obtained from combinatorial and natural product libraries).

One candidate compound is a soluble full-length ubiquitin-specific protease or fragment that competes for substrate binding. Other candidate compounds include mutant ubiquitin proteases or appropriate fragments containing mutations that affect ubiquitin protease function and compete for substrate. Accordingly, a fragment that competes for substrate, for example with a higher affinity, or a fragment that binds substrate but does not hydrolyze the peptide bond, is encompassed by the invention. Other candidate compounds include ubiquitinated protein or protein analog that binds to the protease but is not released or released slowly. Other candidate compounds include analogs of the other natural substrates, such as substrate remnants that bind to but are not released or released more slowly. Further candidate compounds include activators of the proteases such as cytokines, including but not limited to, those disclosed herein.

The invention provides other end points to identify compounds that modulate (stimulate or inhibit) ubiquitin protease activity. The assays typically involve an assay of events in the pathway that indicate ubiquitin protease activity. This can include cellular events that result from deubiquitination, such as cell cycle progression, programmed cell death, growth factor-mediated signal transduction, or any of the cellular processes including, but not limited to, those disclosed herein as resulting from deubiquitination. Specific phenotypes include changes in stress response, DNA replication, receptor internalization, cellular transformation or reversal of transformation, and transcriptional silencing. Assays are based on the multiple cellular functions of deubiquitinating enzymes. These enzymes act at various different levels in the regulation of protein ubiquitination.

A deubiquitinating enzyme can degrade a linear polyubiquitin chain into monomeric ubiquitin molecules. Deubiquitinating enzymes, such as isopeptidase-T, can degrade a branched multiubiquitin chain into monomeric ubiquitin molecules. Deubiquitinating enzymes can remove ubiquitin from a ubiquitin-conjugated target protein. The deubiquitinating enzyme, such as FAF or PA700 isopeptidase, can remove polyubiquitin from a ubiquitinated target protein, and thereby rescue the target from degradation by the 26S proteasome. Deubiquitinating enzymes such as Doa-4 can remove polyubiquitin from proteasome degradation products. The result of all of these is to regulate the cellular pool of free monomeric ubiquitin. Accordingly, assays can be based on detection of any of the products produced by hydrolysis/deubiquitination. Further, the expression of genes that are up- or down-regulated by action of the ubiquitin protease can be assayed. In one embodiment, the regulatory region of such genes can be operably linked to a marker that is easily detectable, such as luciferase. Accordingly, any of the biological or biochemical functions mediated by the ubiquitin protease can be used as an endpoint assay. These include all of the biochemical or biochemical/biological events described herein, in the references cited herein, incorporated by reference for these endpoint assay targets, and other functions known to those of ordinary skill in the art.

Binding and/or activating compounds can also be screened by using chimeric ubiquitin protease proteins in which one or more domains, sites, and the like, as disclosed herein, or parts thereof, can be replaced by their heterologous counterparts derived from other ubiquitin proteases. For example, a recognition or binding region can be used that interacts with different substrate specificity and/or affinity than the native ubiquitin protease. Accordingly, a different set of pathway components is available as an end-point assay for activation. Further, sites that are responsible for developmental, temporal, or tissue specificity can be replaced by heterologous sites such that the protease can be detected under conditions of specific developmental, temporal, or tissue-specific expression.

The ubiquitin protease polypeptides are also useful in competition binding assays in methods designed to discover compounds that interact with the ubiquitin protease. Thus, a compound is exposed to a ubiquitin protease polypeptide under conditions that allow the compound to bind to or to otherwise interact with the polypeptide. Soluble ubiquitin protease polypeptide is also added to the mixture. If the test compound interacts with the soluble ubiquitin protease polypeptide, it decreases the amount of complex formed or activity from the ubiquitin protease target. This type of assay is particularly useful in cases in which compounds are sought that interact with specific regions of the ubiquitin protease. Thus, the

soluble polypeptide that competes with the target ubiquitin protease region is designed to contain peptide sequences corresponding to the region of interest.

Another type of competition-binding assay can be used to discover compounds that interact with specific functional sites. As an example, ubiquitin and a candidate compound can be added to a sample of the ubiquitin protease. Compounds that interact with the ubiquitin protease at the same site as ubiquitin will reduce the amount of complex formed between the ubiquitin protease and ubiquitin. Accordingly, it is possible to discover a compound that specifically prevents interaction between the ubiquitin protease and ubiquitin.

Another example involves adding a candidate compound to a sample of ubiquitin protease and polyubiquitin. A compound that competes with polyubiquitin will reduce the amount of hydrolysis or binding of the polyubiquitin to the ubiquitin protease. Accordingly, compounds can be discovered that directly interact with the ubiquitin protease and compete with polyubiquitin. Such assays can involve any other component that interacts with the ubiquitin protease, such as ubiquitinated substrate protein, ubiquitinated substrate remnants, and cellular components with which the protease interacts such as transcriptional regulatory factors. To perform cell free drug screening assays, it is desirable to immobilize either the ubiquitin protease, or fragment, or its target molecule to facilitate separation of complexes from uncomplexed forms of one or both of the proteins, as well as to accommodate automation of the assay. Techniques for immobilizing proteins on matrices can be used in the drug screening assays.

In one embodiment, a fusion protein can be provided which adds a domain that allows the protein to be bound to a matrix. For example, glutathione-S-transferase/ubiquitin protease fusion proteins can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, Mo.) or glutathione derivatized microtitre plates, which are then combined with the cell lysates (e.g., 35 S-labeled) and the candidate compound, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads are washed to remove any unbound label, and the matrix immobilized and radiolabel determined directly, or in the supernatant after the complexes is dissociated.

Alternatively, the complexes can be dissociated from the matrix, separated by SDS-PAGE, and the level of ubiquitin protease-binding protein found in the bead fraction quantitated from the gel using standard electrophoretic techniques. For example, either the polypeptide or its target molecule can be immobilized utilizing conjugation of biotin and streptavidin using techniques well known in the art. Alternatively, antibodies reactive with the

protein but which do not interfere with binding of the protein to its target molecule can be derivatized to the wells of the plate, and the protein trapped in the wells by antibody conjugation. Preparations of a ubiquitin protease-binding target component, such as ubiquitin, polyubiquitin, ubiquitinated substrate protein, ubiquitinated substrate protein remnant, or ubiquitinated remnant amino acid, and a candidate compound are incubated in the ubiquitin protease-presenting wells and the amount of complex trapped in the well can be quantitated.

Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the ubiquitin protease target molecule, or which are reactive with ubiquitin protease and compete with the target molecule; as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the target molecule.

Modulators of ubiquitin protease activity identified according to these drug screening assays can be used to treat a subject with a disorder mediated by the ubiquitin protease pathway, by treating cells that express the ubiquitin protease.

In one aspect, the invention relates to modulators of the expression of a gene comprising a nucleic acid sequence as set forth in SEQ ID No. 2, SEQ ID No. 6 or SEQ.ID No. 10. Preferably, such modulators are inhibitory nucleic acids, such as antisense oligonucleotides, triple helix DNA, siRNA, ribozymes, RNA aptamers or double or single stranded RNA. It is well within the knowledge of the skilled person to design nucleic acids inhibiting the expression of a gene comprising a nucleic acid sequence as set forth in SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10. In a particularly preferred embodiment the inhibitory nucleic acid is an siRNA. Preferred siRNA molecules are typically between 18 and 30 nucleotides in length, though also greater lengths are suitable to inhibit the expression of the target gene. In a preferred embodiment, the siRNA molecules are between 19 and 25 nucleotides long.

In accordance with the present invention, it has been found that in breast cancer tissue elevated amount of a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ.ID. No.1, SEQ.ID.No.5, or SEQ.ID No. 9 are present. Thus, in one aspect the present invention provides a method for the treatment of breast cancer comprising administering an effective amount of an inhibitory nucleic acid suitable to inhibit the expression of a gene comprising an nucleic acid sequence selected from the group consisting of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10, to a breast cancer patient. In a preferred embodiment, the inhibitory nucleic acid is an siRNA. In another preferred

embodiment, the nucleic acid sequence is SEQ.ID. No.2. In another embodiment, the present invention provides the use of an inbitory nucleic acid, preferably an siRNA, suitable to inhbit the expression of a gene comprising a nucleic acid sequence selected from the group consisting of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10, preferably SEQ ID No. 2, for the manufacture of a medicament for the treatment of breast cancer.

In accordance with the present invention, it has been found that in peripheral blood cells, in particular in lymphoid cells, elevated amount of a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ.ID. No.1, SEQ.ID.No.5, or SEQ.ID No. 9 are present. Thus, in one aspect the present invention provides a method for the treatment of leukemia comprising administering an effective amount of an inhibitory nucleic acid suitable to inhibit the expression of a gene comprising an nucleic acid sequence selected from the group consisting of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10, to a leukemia patient. In a preferred embodiment, the inhibitory nucleic acid is an siRNA. In another preferred embodiment, the nucleic acid sequence is SEQ.ID. No.6. In another embodiment, the present invention provides the use of an inbitory nucleic acid, preferably an siRNA, suitable to inhbit the expression of a gene comprising an nucleic acid sequence selected from the group consisting of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10, preferably SEQ ID No. 6, for the manufacture of a medicament for the treatment of leukemia.

In accordance with another aspect of the present invention, there is provided a pharmaceutical composition comprising an inhibitory nucleic acid, in particular an siRNA, suitable to inhibit the expression of a gene comprising an nucleic acid sequence selected from the group consisting of SEQ ID No. 2, SEQ ID No. 6, SEQ.ID No. 10 and pharmaceutically acceptable carrier.

In one aspect of the invention a method for the diagnosis of diseases involving an ubiquitin-specific protease is provided, comprising measuring the amount of a polynucleotide or polypeptide of the invention in tissue from a human, wherein the presence of an elevated amount of said polynucleotide or polypeptide relative to the amount of said polynucleotide or polypeptide in normal control tissue is diagnostic of said human's suffering from a disease related to an ubiquitin-specific protease. The term "normal" in the context of tissue used for diagnostics methods according to the present invention means tissue from a human not suffering from the disease to be diagnosed. In a preferred embodiment of this aspect, such diagnostic methods are carried out in vitro or ex vivo.

In a preferred aspect of the invention, a method for diagnosis of diseases involving an ubiquitin-specific protease comprises a detection step involving contacting a tissue with an antibody which specifically binds to a polypeptide that comprises the amino acid sequence set forth in any one of SEQ.ID. 1, SEQ ID. No.5 or SEQ ID. No.9 and detecting specific binding of said antibody with a polypeptide in said tissue, wherein detection of specific binding to a polypeptide indicates the presence of a polypeptide that comprises the amino acid set forth in any one of SEQ.ID. 1, SEQ ID. No.5 or SEQ ID. No.9.

In one embodiment of the invention, the cells that are diagnosed are breast cells and the disease involved includes but are not limited to disorders of development of the breast, inflammations, including but not limited to, acute mastitis, periductal mastitis (recurrent subareolar abscess, squamous metaplasia of lactiferous ducts), mammary duct ectasia, fat necrosis, granulomatous mastitis, and pathologies associated with silicone breast implants; fibrocystic changes; proliferative breast disease including, but not limited to, epithelial hyperplasia, sclerosing adenosis, and small duct papillomas; tumors including, but not limited to, stromal tumors such as fibroadenoma, phyllodes tumor, and sarcomas, and epithelial tumors, such as large duct papilloma; carcinoma of the breast including in situ (noninvasive) carcinoma that includes ductal carcinoma in situ (including Paget's disease) and lobular carcinoma in situ, and invasive (infiltrating) carcinoma including, but not limited to, invasive ductal carcinoma, no special type, invasive lobular carcinoma, medullary carcinoma, colloid (mucinous) carcinoma, tubular carcinoma, and invasive papillary carcinoma, and miscellaneous malignant neoplasms. Disorders in the male breast include, but are not limited to, gynecomastia and carcinoma.

In another embodiment of the invention, the cells that are diagnosed are peripheral blood cells, in particular lymphoid cells. Disorders include but are not limited to leukemias.

In yet another embodiment of the invention, the cells that are diagnosed are in the brain and involve disorders of the brain which include but are not limited to disorders involving neurons, and disorders involving glia, such as astrocytes, oligodendrocytes, ependymal cells, and microglia; cerebral edema, raised intracranial pressure and herniation, and hydrocephalus; malformations and developmental diseases, such as neural tube defects, forebrain anomalies, posterior fossa anomalies, and syringomyelia and hydromyelia; perinatal brain injury; cerebrovascular diseases, such as those related to hypoxia, ischemia, and infarction, including hypotension, hypoperfusion, and low-flow states--global cerebral ischemia and focal cerebral ischemia--infarction from obstruction of local blood supply, intracranial hemorrhage, including intracerebral (intraparenchymal) hemorrhage,

subarachnoid hemorrhage and ruptured berry aneurysms, and vascular malformations, hypertensive cerebrovascular disease, including lacunar infarcts, slit hemorrhages, and hypertensive encephalopathy; infections, such as acute meningitis, including acute pyogenic (bacterial) meningitis and acute aseptic (viral) meningitis, acute focal suppurative infections, including brain abscess, subdural empyema, and extradural abscess, chronic bacterial meningoencephalitis, including tuberculosis and mycobacterioses, neurosyphilis, and neuroborreliosis (Lyme disease), viral meningoencephalitis, including arthropod-borne (Arbo) viral encephalitis, Herpes simplex virus Type 1, Herpes simplex virus Type 2, Varicalla - zoster virus (Herpes zoster), cytomegalovirus, poliomyelitis, rabies, and human immunodeficiency virus 1, including HIV-1 meningoencephalitis (subacute encephalitis), vacuolar myelopathy, AIDS-associated myopathy, peripheral neuropathy, and AIDS in children, progressive multifocal leukoencephalopathy, subacute sclerosing panencephalitis, fungal meningoencephalitis, other infectious diseases of the nervous system; transmissible spongiform encephalopathies (prion diseases); demyelinating diseases, including multiple sclerosis, multiple sclerosis variants, acute disseminated encephalomyelitis and acute necrotizing hemorrhagic encephalomyelitis, and other diseases with demyelination; degenerative diseases, such as degenerative diseases affecting the cerebral cortex, including Alzheimer disease and Pick disease, degenerative diseases of basal ganglia and brain stem, including Parkinsonism, idiopathic Parkinson disease (paralysis agitans), progressive supranuclear palsy, corticobasal degenration, multiple system atrophy, including striatonigral degenration, Shy-Drager syndrome, and olivopontocerebellar atrophy, and Huntington disease; spinocerebellar degenerations, including spinocerebellar ataxias, including Friedreich ataxia, and ataxia-telanglectasia, degenerative diseases affecting motor neurons, including amyotrophic lateral sclerosis (motor neuron disease), bulbospinal atrophy (Kennedy syndrome), and spinal muscular atrophy; inborn errors of metabolism, such as leukodystrophies, including Krabbe disease, metachromatic leukodystrophy, adrenoleukodystrophy, Pelizaeus- Merzbacher disease, and Canavan disease, mitochondrial encephalomyopathies, including Leigh disease and other mitochondrial encephalomyopathies; toxic and acquired metabolic diseases, including vitamin deficiencies such as thiamine (vitamin B 1) deficiency and vitamin B 12 deficiency, neurologic sequelae of metabolic disturbances, including hypoglycemia, hyperglycemia, and hepatic encephatopathy, toxic disorders, including carbon monoxide, methanol, ethanol, and radiation, including combined methotrexate and radiation-induced injury; tumors, such as gliomas, including astrocytoma, including fibrillary (diffuse) astrocytoma and glioblastoma

multiforme, pilocytic astrocytoma, pleomorphic xanthoastrocytoma, and brain stem glioma, oligodendroglioma, and ependymoma and related paraventricular mass lesions, neuronal tumors, poorly differentiated neoplasms, including medulloblastoma, other parenchymal tumors, including primary brain lymphoma, germ cell tumors, and pineal parenchymal tumors, meningiomas, metastatic tumors, paraneoplastic syndromes, peripheral nerve sheath tumors, including schwannoma, neurofibroma, and malignant peripheral nerve sheath tumor (malignant schwannoma), and neurocutaneous syndromes (phakomatoses), including neurofibromotosis, including Type 1 neurofibromatosis (NF1) and TYPE 2 neurofibromatosis (NF2), tuberous sclerosis, and Von Hippel-Lindau disease.

This invention also relates to the use of polypeptides of the invention to provide a target for diagnosing a disease or predisposition to disease mediated by the ubiquitin specific proteases, including, but not limited to, diseases involving tissues in which the ubiquitin proteases are expressed as disclosed herein, such as in breast cancer. Accordingly, methods are provided for detecting the presence, or levels of, the ubiquitin protease in a cell, tissue, or organism. The method involves contacting a biological sample with a compound capable of interacting with the ubiquitin protease such that the interaction can be detected. The polypeptides are also useful for treating a disorder characterized by reduced amounts of these components. Thus, increasing or decreasing the activity of the protease is beneficial to treatment. The polypeptides are also useful to provide a target for diagnosing a disease characterized by excessive substrate or reduced levels of substrate. Accordingly, where substrate is excessive, use of the protease polypeptides can provide a diagnostic assay.

Furthermore, for example, proteases having reduced activity can be used to diagnose conditions in which reduced substrate is responsible for the disorder. One agent for detecting ubiquitin protease is an antibody capable of selectively binding to ubiquitin protease. A biological sample includes tissues, cells and biological fluids isolated from a subject, as well as tissues, cells and fluids present within a subject. The ubiquitin protease also provides a target for diagnosing active disease, or predisposition to disease, in a patient having a variant ubiquitin protease. Thus, ubiquitin protease can be isolated from a biological sample and assayed for the presence of a genetic mutation that results in an aberrant protein. This includes amino acid substitution, deletion, insertion, rearrangement, (as the result of aberrant splicing events), and inappropriate post-translational modification. Analytic methods include altered electrophoretic mobility, altered tryptic peptide digest, altered ubiquitin protease activity in cell-based or cell-free assay, alteration in binding to or

hydrolysis of polyubiquitin, binding to ubiquitinated substrate protein or hydrolysis of the ubiquitin from the protein, binding to ubiquitinated protein remnant, including peptide or amino acid, and hydrolysis of the ubiquitin from the remnant, general protein turnover, specific protein turnover, antibody-binding pattern, altered isoelectric point, direct amino acid sequencing, and any other of the known assay techniques useful for detecting mutations in a protein in general or in a ubiquitin protease specifically, including assays discussed herein. In vitro techniques for detection of ubiquitin protease include enzyme linked immunosorbent assays (ELISAs), Western blots, immunoprecipitations and immunofluorescence.

Alternatively, the protein can be detected in vivo in a subject by introducing into the subject a labeled anti- ubiquitin protease antibody. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques. Particularly useful are methods, which detect the allelic variant of the ubiquitin protease expressed in a subject, and methods, which detect fragments of the ubiquitin protease in a sample. The ubiquitin protease polypeptides are also useful in pharmacogenomic analysis. Pharmacogenomics deal with clinically significant hereditary variations in the response to drugs due to altered drug disposition and abnormal action in affected persons. See, e.g., Eichelbaum, M. (1996) Clin. Exp. Pharmacol. Physiol. 23(10-11):983-985, and Linder, M. W. (1997) Clin. Chem. 43(2):254-266. The clinical outcomes of these variations result in severe toxicity of therapeutic drugs in certain individuals or therapeutic failure of drugs in certain individuals as a result of individual variation in metabolism. Thus, the genotype of the individual can determine the way a therapeutic compound acts on the body or the way the body metabolizes the compound. Further, the activity of drug metabolizing enzymes affects both the intensity and duration of drug action. Thus, the pharmacogenomics of the individual permit the selection of effective compounds and effective dosages of such compounds for prophylactic or therapeutic treatment based on the individual's genotype. The discovery of genetic polymorphisms in some drug metabolizing enzymes has explained why some patients do not obtain the expected drug effects, show an exaggerated drug effect, or experience serious toxicity from standard drug dosages. Polymorphisms can be expressed in the phenotype of the extensive metabolizer and the phenotype of the poor metabolizer. Accordingly, genetic polymorphism may lead to allelic protein variants of the ubiquitin protease in which one or more of the ubiquitin protease functions in one population is different from those in another population. The polypeptides thus allow a target to ascertain a genetic predisposition that can affect treatment modality. Thus, in a ubiquitin- based treatment, polymorphism may give rise to

catalytic regions that are more or less active. Accordingly, dosage would necessarily be modified to maximize the therapeutic effect within a given population containing the polymorphism. As an alternative to genotyping, specific polymorphic polypeptides could be identified.

The ubiquitin protease polypeptides are also useful for monitoring therapeutic effects during clinical trials and other treatment. Thus, the therapeutic effectiveness of an agent that is designed to increase or decrease gene expression, protein levels or ubiquitin protease activity can be monitored over the course of treatment using the ubiquitin protease polypeptides as an end-point target. The monitoring can be, for example, as follows: (i) obtaining a pre-administration sample from a subject prior to administration of the agent; (ii) detecting the level of expression or activity of the protein in the pre-administration sample; (iii) obtaining one or more post-administration samples from the subject; (iv) detecting the level of expression or activity of the protein in the post-administration samples; (v) comparing the level of expression or activity of the protein in the pre-administration sample with the protein in the post-administration sample or samples; and (vi) increasing or decreasing the administration of the agent to the subject accordingly.

In another aspect of the invention, methods for treatment include but are not limited to the use of soluble ubiquitin-specific proteases or fragments of the ubiquitin-specific protease protein that compete for substrates including those disclosed herein. These ubiquitin proteases or fragments can have a higher affinity for the target so as to provide effective competition. Stimulation of activity is desirable in situations in which the protein is abnormally downregulated and/or in which increased activity is likely to have a beneficial effect. Likewise, inhibition of activity is desirable in situations in which the protein is abnormally upregulated and/or in which decreased activity is likely to have a beneficial effect. In one example of such a situation, a subject has a disorder characterized by aberrant development or cellular differentiation. In another example, the subject has a proliferative disease (e.g., cancer) or a disorder characterized by an aberrant hematopoietic response. In another example, it is desirable to achieve tissue regeneration in a subject (e.g., where a subject has undergone brain or spinal cord injury and it is desirable to regenerate neuronal tissue in a regulated manner).

In another aspect of the invention, methods for the production of antibodies capable of specifically recognizing one or more differentially expressed gene epitopes are provided. Such antibodies may include, but are not limited to polyclonal antibodies, monoclonal antibodies (mAbs), humanized or chimeric antibodies, single chain antibodies, Fab

fragments, $F(ab')_2$ fragments, fragments produced by a Fab expression library, anti-idiotypic (anti-Id) antibodies, and epitope-binding fragments of any of the above. Such antibodies may be used, for example, in the detection of a fingerprint, target, gene in a biological sample, or, alternatively, as a method for the inhibition of abnormal target gene activity.

For the production of antibodies to a differentially expressed gene, various host animals may be immunized by injection with a differentially expressed gene protein, or a portion thereof. Such host animals may include but are not limited to rabbits, mice, and rats, to name but a few. Various adjuvants may be used to increase the immunological response, depending on the host species, including but not limited to Freund's (complete and incomplete), mineral gels such as aluminum hydroxide, surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanin, dinitrophenol, and potentially useful human adjuvants such as BCG (bacille Calmette-Guerin) and Corynebacterium parvum.

Polyclonal antibodies are heterogeneous populations of antibody molecules derived from the sera of animals immunized with an antigen, such as target gene product, or an antigenic functional derivative thereof. For the production of polyclonal antibodies, host animals such as those described above, may be immunized by injection with differentially expressed gene product supplemented with adjuvants as also described above.

Monoclonal antibodies, which are homogeneous populations of antibodies to a particular antigen, may be obtained by any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to the hybridoma technique of Kohler and Milstein, (1975, Nature 256:495-497; and U.S. Pat. No. 4,376,110), the human B-cell hybridoma technique (Kosbor et al., 1983, Immunology Today 4:72; Cole et al., 1983, Proc. Natl. Acad. Sci. USA 80:2026-2030), and the EBV-hybridoma technique (Cole et al., 1985, Monoclonal Antibodies And Cancer Therapy, Alan R. Liss, Inc., pp. 77-96). Such antibodies may be of any immunoglobulin class including IgG, IgM, IgE, IgA, IgD and any subclass thereof. The hybridoma producing the mAb of this invention may be cultivated in vitro or in vivo. Production of high titers of mAbs in vivo makes this the presently preferred method of production.

In addition, techniques developed for the production of "chimeric antibodies" (Morrison et al., 1984, Proc. Natl. Acad. Sci., 81:6851-6855; Neuberger et al., 1984, Nature, 312:604-608; Takeda et al., 1985, Nature, 314:452-454) by splicing the genes from a mouse antibody molecule of appropriate antigen specificity together with genes from a human antibody molecule of appropriate biological activity can be used. A chimeric antibody is a

molecule in which different portions are derived from different animal species, such as those having a variable or hypervariable region derived from a murine mAb and a human immunoglobulin constant region.

Alternatively, techniques described for the production of single chain antibodies (U.S. Pat. No. 4,946,778; Bird, 1988, Science 242:423-426; Huston et al., 1988, Proc. Natl. Acad. Sci. USA 85:5879-5883; and Ward et al., 1989, Nature 334:544-546) can be adapted to produce differentially expressed gene-single chain antibodies. Single chain antibodies are formed by linking the heavy and light chain fragments of the Fv region via an amino acid bridge, resulting in a single chain polypeptide.

Most preferably, techniques useful for the production of "humanized antibodies" can be adapted to produce antibodies to the polypeptides, fragments, derivatives, and functional equivalents disclosed herein. Such techniques are disclosed in U.S. Patent Nos. 5,932, 448; 5,693,762; 5,693,761; 5,585,089; 5,530,101; 5,910,771; 5,569,825; 5,625,126; 5,633,425; 5,789,650; 5,545,580; 5,661,016; and 5,770,429, the disclosures of all of which are incorporated by reference herein in their entirety.

Antibody fragments which recognize specific epitopes may be generated by known techniques. For example, such fragments include but are not limited to: the $F(ab')_2$ fragments which can be produced by pepsin digestion of the antibody molecule and the Fab fragments which can be generated by reducing the disulfide bridges of the $F(ab')_2$ fragments. Alternatively, Fab expression libraries may be constructed (Huse et al., 1989, Science, 246:1275-1281) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity.

Particularly preferred, for ease of detection, is the sandwich assay, of which a number of variations exist, all of which are intended to be encompassed by the present invention.

For example, in a typical forward assay, unlabeled antibody is immobilized on a solid substrate and the sample to be tested brought into contact with the bound molecule. After a suitable period of incubation, for a period of time sufficient to allow formation of an antibody-antigen binary complex. At this point, a second antibody, labeled with a reporter molecule capable of inducing a detectable signal, is then added and incubated, allowing time sufficient for the formation of a ternary complex of antibody-antigen-labeled antibody. Any unreacted material is washed away, and the presence of the antigen is determined by observation of a signal, or may be quantitated by comparing with a control sample containing known amounts of antigen. Variations on the forward assay include the simultaneous assay, in which both

sample and antibody are added simultaneously to the bound antibody, or a reverse assay in which the labeled antibody and sample to be tested are first combined, incubated and added to the unlabeled surface bound antibody. These techniques are well known to those skilled in the art, and the possibility of minor variations will be readily apparent. As used herein, "sandwich assay" is intended to encompass all variations on the basic two-site technique.

The most commonly used reporter molecules in this type of assay are either enzymes, fluorophore- or radionuclide-containing molecules. In the case of an enzyme immunoassay an enzyme is conjugated to the second antibody, usually by means of glutaraldehyde or periodate. As will be readily recognized, however, a wide variety of different ligation techniques exist, which ae well-known to the skilled artisan. Commonly used enzymes include horseradish peroxidase, glucose oxidase, beta-galactosidase and alkaline phosphatase, among others. The substrates to be used with the specific enzymes are generally chosen for the production, upon hydrolysis by the corresponding enzyme, of a detectable color change. Alternately, fluorescent compounds, such as fluorescein and rhodamine, may be chemically coupled to antibodies without altering their binding capacity. When activated by illumination with light of a particular wavelength, the fluorochrome-labeled antibody absorbs the light energy, inducing a state of excitability in the molecule, followed by emission of the light at a characteristic longer wavelength. The emission appears as a characteristic color visually detectable with a light microscope. Immunofluorescence and EIA techniques are both very well established in the art and are particularly preferred for the present method. However, other reporter molecules, such as radioisotopes, chemiluminescent or bioluminescent molecules may also be employed. It will be readily apparent to the skilled artisan how to vary the procedure to suit the required use.

Thus in another aspect, the present invention relates to a diagnostic kit which comprises at least one of the following components: (a) an oligonucleotide suitable to detect a nucleic acid comprising a nucleotide sequence or part of a nucleotide sequence as set forth in SEQ ID No.2, SEQ ID. No.6 or SEQ ID No.10, (b) an antibody suitable to detect a polypeptide comprising an amino acid sequence or part of an amino acid sequence as set forth in SEQ ID. No. 1, SEQ ID No. 5 or SEQ ID. No.9, (c) instruction for using the kit.

The nucleotide sequences of the present invention are also valuable for chromosome localization. The sequence is specifically targeted to, and can hybridize with, a particular location on an individual human chromosome. The mapping of relevant sequences to chromosomes according to the present invention is an important first step in correlating those sequences with gene associated disease. Once a sequence has been mapped to a

precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. Such data are found in, for example, V. McKusick, Mendelian Inheritance in Man (available on-line through Johns Hopkins University Welch Medical Library). The relationship between genes and diseases that have been mapped to the same chromosomal region are then identified through linkage analysis (coinheritance of physically adjacent genes).

The differences in the cDNA or genomic sequence between affected and unaffected individuals can also be determined. If a mutation is observed in some or all of the affected individuals but not in any normal individuals, then the mutation is likely to be the causative agent of the disease.

An additional aspect of the invention relates to the administration of a pharmaceutical composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic effects discussed above. Such pharmaceutical compositions may consist of antibodies, mimetics, agonists, antagonists, or inhibitors of the ubiquitin-specific proteases of the present invention. The compositions may be administered alone or in combination with at least one other agent, such as stabilizing compound, which may be administered in any sterile, biocompatible pharmaceutical carrier, including, but not limited to, saline, buffered saline, dextrose, and water. The compositions may be administered to a patient alone, or in combination with other agents, drugs or hormones.

The pharmaceutical compositions encompassed by the invention may be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-articular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

In addition to the active ingredients, these pharmaceutical compositions may contain suitable pharmaceutically-acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, Pa.).

Pharmaceutical compositions for oral administration can be formulated using pharmaceutically acceptable carriers well known in the art in dosages suitable for oral administration. Such carriers enable the pharmaceutical compositions to be formulated as

tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, suspensions, and the like, for ingestion by the patient.

Pharmaceutical preparations for oral use can be obtained through combination of active compounds with solid excipient, optionally grinding a resulting mixture, and processing the mixture of granules, after adding suitable auxiliaries, if desired, to obtain tablets or dragee cores. Suitable excipients are carbohydrate or protein fillers, such as sugars, including lactose, sucrose, mannitol, or sorbitol; starch from corn, wheat, rice, potato, or other plants; cellulose, such as methyl cellulose, hydroxypropylmethyl-cellulose, or sodium carboxymethylcellulose; gums including arabic and tragacanth; and proteins such as gelatin and collagen. If desired, disintegrating or solubilizing agents may be added, such as the cross-linked polyvinyl pyrrolidone, agar, alginic acid, or a salt thereof, such as sodium alginate.

Dragee cores may be used in conjunction with suitable coatings, such as concentrated sugar solutions, which may also contain gum arabic, talc, polyvinylpyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures. Dyestuffs or pigments may be added to the tablets or dragee coatings for product identification or to characterize the quantity of active compound, i.e., dosage.

Pharmaceutical preparations which can be used orally include push-fit capsules made of gelatin, as well as soft, sealed capsules made of gelatin and a coating, such as glycerol or sorbitol. Push-fit capsules can contain active ingredients mixed with a filler or binders, such as lactose or starches, lubricants, such as talc or magnesium stearate, and, optionally, stabilizers. In soft capsules, the active compounds may be dissolved or suspended in suitable liquids, such as fatty oils, liquid, or liquid polyethylene glycol with or without stabilizers.

Pharmaceutical formulations suitable for parenteral administration may be formulated m aqueous solutions, preferably in physiologically compatible buffers such as Hanks' solution, Ringer's solution, or physiologically buffered saline. Aqueous injection suspensions may contain substances which increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Additionally, suspensions of the active compounds may be prepared as appropriate oily injection suspensions. Suitable lipophilic solvents or vehicles include fatty oils such as sesame oil, or synthetic fatty acid esters, such as ethyl oleate or triglycerides, or liposomes. Non-lipid polycationic amino polymers may also be used for delivery. Optionally, the suspension may also contain suitable stabilizers or

agents which increase the solubility of the compounds to allow for the preparation of highly concentrated solutions.

For topical or nasal administration, penetrants appropriate to the particular barrier to be permeated are used in the formulation. Such penetrants are generally known in the art.

The pharmaceutical compositions of the present invention may be manufactured in a manner that is known in the art, e.g., by means of conventional mixing, dissolving, granulating, dragee-making, levigating, emulsifying, encapsulating, entrapping, or lyophilizing processes.

The pharmaceutical composition may be provided as a salt and can be formed with many acids, including but not limited to, hydrochloric, sulfuric, acetic, lactic, tartaric, malic, succinic, etc. Salts tend to be more soluble in aqueous or other protonic solvents than are the corresponding free base forms. In other cases, the preferred preparation may be a lyophilized powder which may contain any or all of the following: 1-50 mM histidine, 0. 1%-2% sucrose, and 2-7% mannitol, at a pH range of 4.5 to 5.5, that is combined with buffer prior to use.

After pharmaceutical compositions have been prepared, they can be placed in an appropriate container and labeled for treatment of an indicated condition. For administration labeling would include amount, frequency, and method of administration.

Pharmaceutical compositions suitable for use in the invention include compositions wherein the active ingredients are contained in an effective amount to achieve the intended purpose. The determination of an effective dose is well within the capability of those skilled in the art.

For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays, e.g., of neoplastic cells, or in animal models, usually mice, rabbits, dogs, or pigs. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient, fragments thereof, antibodies, agonists, antagonists or inhibitors of the ubiquitin-specific protease. which ameliorates the symptoms or condition. Therapeutic efficacy and toxicity may be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., ED50 (the dose therapeutically effective in 50% of the population) and LD50 (the dose lethal to 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index, and it can be expressed as the ratio, LD50/ED50. Pharmaceutical

compositions which exhibit large therapeutic indices are preferred. The data obtained from cell culture assays and animal studies is used in formulating a range of dosage for human use. The dosage contained in such compositions is preferably within a range of circulating concentrations that include the ED50 with little or no toxicity. The dosage varies within this range depending upon the dosage form employed, sensitivity of the patient, and the route of administration.

The exact dosage will be determined by the practitioner, in light of factors related to the subject that requires treatment. Dosage and administration are adjusted to provide sufficient levels of the active moiety or to maintain the desired effect. Factors which may be taken into account include the severity of the disease state, general health of the subject, age, weight, and gender of the subject, diet, time and frequency of administration, drug combination(s), reaction sensitivities, and tolerance/response to therapy. Long-acting pharmaceutical compositions may be administered every 3 to 4 days, every week, or once every two weeks depending on half-life and clearance rate of the particular formulation.

Normal dosage amounts may vary from 0.1 to 100,000 micrograms, up to a total dose of about 1 g, depending upon the route of administration. Guidance as to particular dosages and methods of delivery is provided in the literature and generally available to practitioners in the art. Those skilled in the art will employ different formulations for nucleotides than for proteins or their inhibitors. Similarly, delivery of polynucleotides or polypeptides will be specific to particular cells, conditions, locations, etc. Pharmaceutical formulations suitable for oral administration of proteins are described, e.g., in U.S. Patents 5,008,114; 5,505,962; 5,641,515; 5,681,811; 5,700,486; 5,766,633; 5,792,451; 5,853,748; 5,972,387; 5,976,569; and 6,051,561.

The following examples and tables illustrate the present invention, without in any way limiting the scope thereof.

Tables:

Table 1(a)-USP N01: novel splice variant -amino acid sequence

```
LOCUS            USP_N01     3370 aa     PRT
Accession   BAA25496
GeneSeq     AAU82706
ORIGIN   human

Splice form characterized by 4 insertions:
Pos 1 - Pos 11
Pos 267 - Pos 300
Pos 361 - Pos 384
Pos 1243 - Pos 1437

  1 MRRKNSYYVW QKIFQIQFPL YTAYKHNTHP TIEDISTQES NILGAFCDMN DVEVPLHLLR
 61 YVCLFCGKNG LSLMKDCFEY GTPETLPFLI AHAFITVVSN IRIWLHIPAV MQHIIPFRTY
121 VIRYLCKLSD QELRQSAARN MADLMWSTVK EPLDTTLCFD KESLDLAFKY FMSPTLTMRL
181 AGLSQITNQL HTFNDVCNNE SLVSDTETSI AKELADWLIS NNVVEHIFGP NLHIEIIKQC
```

241 QVILNFLAAE GRLSTQHIDC IWAAAQRKKS LEEQLHHLGH LQLVLKAVII AIHIKVEVVT
301 EPSVHTEQTL YLASMLIKAL WNNALAAKAQ LSKQSSFASL LNTNIPIGNK KEEEELRRTA
361 LKWMSNLLIE PNMCNNDFQT QKESMQGSSD ETANSGEDGS SGPGSSSGHS DGSSNEVNSS
421 HASQSAGSPG SEVQSEDIAD IEALKEEDED DDHGHNPPKS SCGTDLRNRK LESQAGICLG
481 DSQGTSERNG TSSGTGKDLV FNTESLPSVD NRMRMLDACS HSEDPEHDIS GEMNATHIAQ
541 GSQESCITRT GDFLGETIGN ELFNCRQFIG PQHHHHHHHH HHHHDGHMVD DMLSADDVSC
601 SSSQVSAKSE KNMADFDGEE SGCEEELVQI NSHAELTSHL QQHLPNLASI YHEHLSQGPV
661 VHKHQFNSNA VTDINLDNVC KKGNTLLWDI VQDEDAVNLS EGLINEAEKL LCSLVCWFTD
721 RQIRMRFIEG CLENLGNNRS VVISLRLLPK LFGTFQQFGS SYDTHWITMW AEKELNMMKL
781 FFDNLVYYIQ TVREGRQKHA LYSHSAEVQV RLQFLTCVFS TLGSPDHFRL SLEQVDILWH
841 CLVEDSECYD DALHWFLNQV RSKDQHAMGM ETYKHLFLEK MPQLKPETIS MTGLNLFQHL
901 CNLARLATSA YDGCSNSELC GMDQFWGIAL RAQSGDVSRA AIQYINSYYI NGKTGLEKEQ
961 EFISKCMESL MIASSSLEQE SHSSLMVIER GLLMLKTHLE AFRRRFAYHL RQWQIEGTGI
1021 SSHLKALSDK QSLPLRVVCQ PAGLPDKMTI EMYPSDQVAD LRAEVTHWYE NLQKEQINQQ
1081 AQLQEFGQSN RKGEFPGGLM GPVRMISSGH ELTTDYDEKA LHELGFKDMQ MVFVSLGAPR
1141 RERKGEGVQL PASCLPPPQK DNIPMLLLLQ EPHLTTLFDL LEMLASFKPP SGKVAVDDSE
1201 SLRCEELHLH AENLSRRVWE LLMLLPTCPN MLMAFQNISD EQSNDGFNWK ELLKIKSAHK
1261 LLYALEIIEA LGKPNRRIRR ESTGSYSDLY PDSDDSSEDQ VENSKNSWSC KFVAAGGLQQ
1321 LLEIFNSGIL EPKEQESWTV WQLDCLACLL KLICQFAVDP SDLDLAYHDV FAWSGIAESH
1381 RKRTWPGKSR KAAGDHAKGL HIPRLTEVFL VLVQGTSLIQ RLMSVAYTYD NLAPRVLKAQ
1441 SDHRSRHEVS HYSMWLLVSW AHCCSLVKSS LADSDHLQDW LKKLTLLIPE TAVRHESCSG
1501 LYKLSLSGLD GGDSINRSFL LLAASTLLKF LPDAQALKPI RIDDYEEEPI LKPGCKEYFW
1561 LLCKLVDNIH IKDASQTTLL DLDALARHLA DCIRSREILD HQDGNVEDDG LTGLLRLATS
1621 VVKHKPPFKF SREGQEFLRD IFNLLFLLPS LKDRQQPKCK SHSSRAAAYD LLVEMVKGSV
1681 ENYRLIHNWV MAQHMQSHAP YKWDYWPHED VRAECRFVGL TNLGATCYLA STIQQLYMIP
1741 EARQAVFTAK YSEDMKHKTT LLELQKMFTY LMESECKAYN PRPFCKTYTM DKQPLNTGEQ
1801 KDMTEFFTDL ITKIEEMSPE LKNTVKSLFG GVITNNVVSL DCEHVSQTAE EFYTVRCQVA
1861 DMKNIYESLD EVTIKDTLEG DNMYTCSHCG KKVRAEKRAC FKKLPRILSF NTMRYTFNMV
1921 TMMKEKVNTH FSFPLRLDMT PYTEDFLMGK SERKEGFKEV SDHSKDSESY EYDLIGVTVH
1981 TGTADGGHYY SFIRDIVNPH AYKNNKWYLF NDAEVKPFDS AQLASECFGG EMTTKTYDSV
2041 TDKFMDFSFE KTHSAYMLFY KRMEPEEENG REYKFDVSSE LLEWIWHDNM QFLQDKNIFE
2101 HTYFGFMWQL CSCIPSTLPD PKAVSLMTAK LSTSFVLETF IHSKEKPTML QWIELLTKQF
2161 NNSQAACEWF LDRMADDDWW PMQILIKCPN QIVRQMFQRL CIHVIQRLRP VHAHLYLQPG
2221 MEDGSDDMDT SVEDIGGRSC VTRFVRTLLL IMEHGVKPHS KHLTEYFAFL YEFAKMGEEE
2281 SQFLLSLQAI STMVHFYMGT KGPENPQVEV LSEEEGEEEE EEEDILSLAE EKYRPAALEK
2341 MIALVALLVE QSRSERHLTL SQTDMAALTG GKGFPFLFQH IRDGINIRQT CNLIFSLCRY
2401 NNRLAEHIVS MLFTSIAKLT PEAANPFFKL LTMLMEFAGG PPGMPPFASY ILQRIWEVIE
2461 YNPSQCLDWL AVQTPRNKLA HSWVLQNMEN WVERFLLAHN YPRVRTSAAY LLVSLIPSNS
2521 FRQMFRSTRS LHIPTRDLPL SPDTTVVLHQ VYNVLLGLLS RAKLYVDAAV HGTTKLVPYF
2581 SFMTYCLISK TEKLMFSTYF MDLWNLFQPK LSEPAIATNH NKQALLSFWY NVCADCPENI
2641 RLIVQNPVVT KNIAFNYILA DHDDQDVVLF NRGMLPAYYG ILRLCCEQSP AFTRQLASHQ
2701 NIQWAFKNLT PHASQYPGAV EELFNLMQLF IAQRPDMREE ELEDIKQFKK TTISCYLRCL
2761 DGRSCWTTLI SAFRILLESD EDRLLVVFNR GLILMTESFN TLHMMYHEAT ACHVTGDLVE
2821 LLSIFLSVLK STRPYLQRKD VKQALIQWQE RIEFAHKLLT LLNSYSPPEL RNACIDVLKE
2881 LVLLSPHDFL HTLVPFLQHN HCTYHHSNIP MSLGPYFPCR ENIKLIGGKS NIRPPRPELN
2941 MCLLPTMVET SKGKDDVYDR MLLDYFFSYH QFIHLLCRVA INCEKFTETL VKLSVLVAYE
3001 GLPLHLALFP KLWTELCQTQ SAMSKNCIKL LCEDPVFAEY IKCILMDERT FLNNNIVYTF
3061 MTHFLLKVQS QVFSEANCAN LISTLITNLI SQYQNLQSDF SNRVEISKAS ASLNGDLRAL
3121 ALLLSVHTPK QLNPALIPTL QELLSKCRTC LQQRNSLQEQ EAKERKTKDD EGATPIKRRR
3181 VSSDEEHTVD SCISDMKTET REVLTPTSTS DNETRDSSII DPGTEQDLPS PENSSVKEYR
3241 MEVPSSFSED MSNIRSQHAE EQSNNGRYDD CKEFKDLHCS KDSTLAEEES EFPSTSISAV
3301 LSDLADLRSC DGQALPSQDP EVALSLSCGH SRGLFSHMQQ HDILDTLCRT IESTIHVVTR
3361 ISGKGNQAAS

Table 1(b):-USP N01: novel splice variant -nucleotide sequence

```
   1 atgagaagga aaaactctta ctatgtgtgg caaaaaattt ttcaaattca gtttccctta
  61 tatactgctt acaagcataa tactcaccct actattgagg atatatcaac tcaagaaagt
 121 aacatattag gggcattctg tgatatgaat gatgtagaag taccattgca tttgcttcgt
 181 tatgtatgtt tgttttgtgg gaaaaatggc ctttctctca tgaaggattg ctttgaatat
 241 ggaactcctg aaactttgcc atttcttata gcacatgcgt ttattacagt tgtgtctaat
 301 attagaatat ggctacatat tcccgctgtc atgcagcaca ttataccttt taggacctat
 361 gttattaggt atttatgcaa gctctcggat caggagttac gacagagtgc agctcgtaac
 421 atggctgact taatgtggag cacagtcaaa gaaccattgg atacaacatt atgctttgat
 481 aaagaaagcc tagatcttgc atttaagtac tttatgtcac ctactttgac tatgaggttg
 541 gctggattga gtcagataac aaatcaactc catacettca atgatgtgtg caataatgaa
 601 tcattagtat cggacacaga aacgtccatt gcaaaagaac ttgcagactg gcttattagc
 661 aacaatgtgg tggagcatat atttggacca aatttacata ttgagattat caaacagtgc
 721 caagtgattt tgaatttttt ggcagcagaa gggcgactga gtactcaaca tattgactgt
 781 atttgggctg cagcacagag gaagaagagc ttagaagaac agctccatca ccttggtcac
 841 ctgcagctag tcctcaaagc agtgataata gcgatacaca tcaaagtgga ggtagtgaca
 901 ttgaaatgga tgagcaactt attaatagaa ccaaacatgt gcaacaacga ctttcagaca
 961 cagaaggaat ccatgcaggg aagttctgac gaaactgcca acagtggtga agatggaagc
1021 agtggtcctg gtagcagtag tgggcatagt gatggatcta gcaatgaggt taattctagc
1081 cacgcaagcc agtcagctgg gagccctggc agtgaggtac agtcagaaga cattgcagat
1141 attgaagccc tcaaagagga agatgaagac gatgatcatg gtcataatcc tcccaaaagc
1201 agttgtggta cagatcttcg gaatagaaag ttagagagtc aagcaggcat ttgcctgggg
1261 gactcccaag gcacgtcaga aagaaatggg acaagcagcg gaacaggaaa ggacctggtt
1321 tttaacactg aatcattgcc atcagtagat aatcgaatgc gaatgctgga tgcttgttca
1381 cactctgaag acccagaaca tgatatttca ggggaaatga atgctactca tatagcacaa
1441 gggtctcagg agtcttgtat cacacgaact ggggacttcc ttggggagac tattgggaat
1501 gaattattta attgtcgaca atttattggt ccacagcatc accaccacca ccaccaccat
1561 caccaccacc acgatgggca tatggttgat gatatgctaa gtgcagatga tgtcagttgt
1621 agtagctccc aggttagtgc aaaatcagaa aaaaatatgg ctgattttga tggtgaagaa
1681 tctggatgtg aagaggagct agttcagatt aattcacatg cggaactgac atctcacctc
1741 caacaacatc ttcccaattt agcttccatt taccatgaac atcttagtca aggacctgta
1801 gttcataaac atcaattcaa cagtaatgct gttacagaca ttaatttgga taatgtttgc
1861 aagaaaggaa atactttgtt gtgggatata gtccaagatg aagatgcagt taatctttct
1921 gaaggattaa taaatgaagc agagaaactt ctttgttcgt tagtatgttg gtttacagat
1981 agacaaattc gaatgagatt cattgaaggt tgccttgaaa acttgggaaa caacagatca
2041 gtagtaattt cacttcgtct tcttccaaaa ctatttggta cttttcagca gtttgggagc
2101 agttacgata cacactggat aacaatgtgg gcagaaaaag aactgaacat gatgaagctt
2161 ttctttgata atttggtata ctacattcaa actgtgagag aaggaagaca aaaacatgca
2221 ctgtacagcc atagtgctga agttcaagtt cgtcttcaat tcttgacttg tgtattttca
2281 actctgggat cacctgatca tttcaggtta agtttagagc aagttgacat cttatggcat
2341 tgtttagtag aagattctga atgttatgat gatgcactcc attggttttt aaatcaagtt
2401 cgaagtaaag atcaacatgc tatgggtatg gaaacctaca aacatctttt cctggagaag
2461 atgccccagc taaaacctga aacaattagc atgactggct taaacctgtt tcagcatctc
2521 tgtaacttgg ctcgattggc taccagtgcc tatgatggtt gttcaaattc tgagctgtgt
2581 ggtatggacc aatttggggg cattgcttta agagcacaat ctggtgatgt cagtcgagca
2641 gctatccagt atattaactc ctattatatt aatggtaaaa caggtttgga gaaggagcaa
2701 gaatttatta gtaagtgcat ggagagtctt atgatagctt ctagcagtct tgaacaggaa
2761 tcacactcaa gtctcatggt tatagaaaga ggactcctta tgctgaagac acatctggaa
2821 gcgtttagga gaaggtttgc atatcatctg agacagtggc aaattgaagg cactggtatt
2881 agtagtcatt tgaaagcact gagtgacaaa cagtctctgc cgctaagggt tgtatgccag
2941 ccagctggac ttcctgacaa gatgactatt gaaatgtatc ctagtgacca ggtagcagat
3001 cttagggctg aagtaactca ttggtatgaa aatttacaga aagaacaaat aaatcaacaa
3061 gctcagcttc aggagtttgg tcaaagcaac cgaaaaggag agtttcctgg aggcctcatg
```

3121 ggacctgtca ggatgatttc atctggacac gagttaacaa cagattatga tgaaaaagca
3181 cttcatgagc ttggttttaa ggatatgcag atggtatttg tatctttggg tgcaccaagg
3241 agagagcgga aaggggaagg tgttcagctg ccagcatctt gcctcccacc ccctcagaag
3301 gacaacattc caatgctttt gcttttacaa gagcctcatt taactactct ttttgattta
3361 ttagagatgc ttgcatcatt taaaccaccc tcaggaaaag tggcagtgga tgatagtgag
3421 agcttacgat gtgaagaact tcatcttcat gcagaaaatc tgtctaggcg ggtctgggag
3481 ctactgatgc ttcttcctac atgtcctaat atgttgatgg cattccagaa tatctcagat
3541 gagcagagta atgatggatt taattggaaa gaacttctca aaattaagag cgcccacaag
3601 ctattgtatg ctctggaaat tattgaagca ctgggaaaac ctaatagaag aataaggagg
3661 gagtctacgg gaagttacag tgatctttat ccagattcag atgattcaag tgaggatcaa
3721 gtggaaaata gtaaaaattc ctggagttgc aagtttgttg ctgctggagg gcttcaacag
3781 ttattagaaa tttttaattc tggaattcta gagcctaaag agcaggaatc atggactgtg
3841 tggcagctag actgtcttgc ttgcttgctg aagttaatat gccagtttgc agtagatcca
3901 tccgatttgg atttagctta tcatgatgtc tttgcctggt ctggtatagc ggaaagccat
3961 aggaaaagaa cctggcctgg caaatcaagg aaggctgctg gtgatcatgc taagggtctt
4021 catataccac gattaacaga ggtatttctt gttcttgtcc aaggaaccag tttgattcag
4081 cgacttatgt ctgttgctta tacgtatgat aatctggctc ctagagtttt aaaagctcag
4141 tctgatcaca ggtctagaca tgaagtttca cattattcaa tgtggctctt ggtgagttgg
4201 gctcattgct gttctttagt gaaatctagc cttgctgata gcgatcattt acaagattgg
4261 ctaaagaaat tgactctcct tattcctgag actgcagttc gtcatgaatc atgcagtggt
4321 ctctataagt tatccctgtc agggctggat ggaggagact caatcaatcg ttcttttctg
4381 ctattggctg cctcaacatt attgaaattt cttcctgatg ctcaagcact caaacctatt
4441 aggatagatg attatgagga agaaccaata ttaaaaccag gatgtaaaga gtatttttgg
4501 ttgttatgca aattagttga caacatacat ataaaggacg ctagtcagac aacgctcctc
4561 gacttagatg ccttggcaag acatttggct gactgtattc gaagtaggga gatccttgat
4621 catcaggatg gtaatgtaga agatgatggg cttacaggac tcctaaggct tgcaacaagt
4681 gttgttaaac acaaaccacc ctttaaattt tcaagggaag gacaggaatt tttgagagat
4741 atcttcaatc tcctgttttt gttgccaagt ctaaaggacc gacaacagcc aaagtgcaaa
4801 tcacattctt caagagctgc cgcttacgat ttgttagtag agatggtaaa ggggtctgtt
4861 gagaactaca ggctaataca caactgggtt atggcacaac acatgcagtc ccatgcacct
4921 tataaatggg attactggcc tcatgaagat gtccgtgctg aatgtagatt tgttggcctt
4981 actaaccttg gagctacttg ttacttagct tctactattc agcaacttta tatgatacct
5041 gaggcaagac aggctgtctt cactgccaag tattcagagg atatgaagca caagaccact
5101 cttctggagc ttcagaaaat gtttacatat ttaatggaga gtgaatgcaa agcatataat
5161 cctagacctt tctgtaaaac atacaccatg gataagcagc ctctgaatac tggggaacag
5221 aaagatatga cagagttttt tactgatcta attaccaaaa tcgaagaaat gtctcccgaa
5281 ctgaaaaata ccgtcaaaag tttatttgga ggtgtaatta caaacaatgt tgtatccttg
5341 gattgtgaac atgttagtca aactgctgaa gagttttata ctgtgaggtg ccaagtggct
5401 gatatgaaga acatttatga atctcttgat gaagttacta taaaagacac tttggaaggt
5461 gataacatgt atacttgttc tcattgtggg aagaaagtac gagctgaaaa aagggcatgt
5521 tttaagaaat tgcctcgcat tttgagtttc aatactatga gatacacatt taatatggtc
5581 acgatgatga aagagaaagt gaatacacac ttttccttcc cattacgttt ggacatgacg
5641 ccctatacag aagattttct tatgggaaag agtgagagga aagaaggttt taaagaagtc
5701 agtgatcatt caaaagactc agagagctat gaatatgact tgataggagt gactgttcac
5761 acaggaacgg cagatggtgg acactattat agctttatca gagatatagt aaatccccat
5821 gcttataaaa acaataaatg gtatcttttt aatgatgctg aggtaaaacc ttttgattct
5881 gctcaacttg catctgaatg ttttggtgga gagatgacga ccaagaccta tgattctgtt
5941 acagataaat ttatggactt ctcttttgaa aagacacaca gtgcatatat gctgttttac
6001 aaacgcatgg aaccagagga agaaaatggc agagaataca aatttgatgt ttcgtcagag
6061 ttactagagt ggatttggca tgataacatg cagtttcttc aagacaaaaa catttttgaa
6121 catacatatt ttggatttat gtggcaattg tgtagttgta ttcccagtac attaccagat
6181 cctaaagctg tgtccttaat gacagcaaag ttaagcactt cctttgtcct agagacattt
6241 attcattcta aagaaaagcc cacgatgctt cagtggattg aactgttgac gaaacagttt
6301 aataatagtc aggcagcttg tgagtggttt ttagatcgta tggctgatga cgactggtgg

6361 ccaatgcaga tactaattaa gtgccctaat caaattgtga gacagatgtt tcagcgtttg
6421 tgtatccatg tgattcagag gctgagacct gtgcatgctc atctctattt gcagccagga
6481 atggaagatg ggtcagatga tatggatacc tcagtagaag atattggtgg tcgttcatgt
6541 gtcactcgct ttgtgagaac cctgttatta attatggaac atggtgtaaa acctcacagt
6601 aaacatctta cagagtattt tgccttcctt tacgaatttg caaaaatggg tgaagaagag
6661 agccaatttt tgctttcatt gcaagctata tctacaatgg tacattttta catgggaaca
6721 aaaggacctg aaaatcctca agttgaagtg ttatcagagg aagaagggga agaagaagag
6781 gaggaagaag atatcctctc tctggcagaa gaaaaataca ggccagctgc ccttgaaaag
6841 atgatagctt tagttgctct tttggttgaa cagtctcgat cagaaaggca tttgacatta
6901 tcacagactg acatggcagc attaacagga ggaaagggat ttcccttctt gtttcaacat
6961 attcgtgatg gcatcaatat aagacaaact tgtaatctga ttttcagcct gtgtcgatac
7021 aataatcgac ttgcagaaca tattgtatct atgcttttca catcaatagc aaagttgact
7081 cctgaggcag ccaatccttt ctttaagttg ttgactatgc taatggagtt tgctggtgga
7141 cctccaggaa tgcctcccct tgcatcttat attctgcaga ggatatggga ggtgattgaa
7201 tacaatcctt ctcagtgtct agattggttg gcagtgcaga caccccgaaa taaactggca
7261 cacagctggg tcttacagaa tatggaaaac tgggtcgagc ggtttctttt ggctcacaat
7321 tatcctagag tgaggacttc tgcagcttat cttctggtgt cccttatacc aagcaattca
7381 ttccgtcaga tgttccggtc aacaaggtct ttgcacatcc caacccgtga ccttccactc
7441 agtccagaca caacagtagt cctacatcag gtctacaacg tgctccttgg tttgctctca
7501 agagccaaac tttatgttga tgctgctgtt catggcacta caaagctagt gccctatttt
7561 agctttatga cttactgttt aatttccaaa actgagaagc tgatgttttc cacatatttc
7621 atggatttgt ggaacctttt ccagcctaaa ctttctgagc cagcaatagc tacaaatcac
7681 aataaacagg ctttgctttc attttggtac aatgtctgtg ctgactgtcc agagaatatc
7741 cgccttattg ttcagaaccc agtggtaacc aagaacattg ccttcaatta catccttgct
7801 gaccatgatg atcaggatgt ggtgcttttt aaccgtggga tgctgccagc gtactatggc
7861 attctgaggc tctgctgtga gcagtctcct gcattcacac gacaactggc ttctcaccag
7921 aacatccagt gggcctttaa gaatcttaca ccacatgcca gccaataccc tggagcagta
7981 gaagaactgt ttaacctgat gcagctgttt atagctcaga ggccagatat gagagaagaa
8041 gaattagaag atattaaaca gttcaagaaa acaaccataa gttgttactt acgttgctta
8101 gatggccgct cctgctggac tactttaata agtgccttca gaatactatt agaatctgat
8161 gaagacagac ttcttgttgt atttaatcga ggattgattc taatgacaga gtctttcaac
8221 actttgcaca tgatgtatca cgaagctaca gcttgccatg tgactggaga tttagtagaa
8281 cttctgtcaa tatttctttc ggttttgaag tctacacgcc cttatcttca gagaaaagat
8341 gtgaaacaag cattaatcca gtggcaggag cgaattgaat ttgcccataa actgttaact
8401 cttcttaatt cctatagtcc tccagaactt agaaatgcct gtatagatgt cctcaaggaa
8461 cttgtacttt tgagtcccca tgattttctt catactctgg ttccctttct acaacacaac
8521 cattgtactt accatcacag taatatacca atgtctcttg gaccttattt cccttgtcga
8581 gaaaatatca agctaatagg agggaaaagc aatattcggc ctccgcgccc tgaactcaat
8641 atgtgcctct tgcccacaat ggtggaaacc agtaagggca aagatgacgt ttatgatcgt
8701 atgctgctag actacttctt ttcttatcat cagttcatcc atctattatg ccgagttgca
8761 atcaactgtg aaaaatttac tgaaacatta gttaagctga gtgtcctagt tgcctatgaa
8821 ggtttgccac ttcatcttgc actgttcccc aaactttgga ctgagctatg ccagactcag
8881 tctgctatgt caaaaaactg catcaagctt ttgtgtgaag atcctgtttt cgcagaatat
8941 attaaatgta tcctaatgga tgaaagaact tttttaaaca acaacattgt ctacacgttc
9001 atgacacatt tccttctaaa ggttcaaagt caagtgtttt ctgaagcaaa ctgtgccaat
9061 ttgatcagca ctcttattac aaacttgata agccagtatc agaacctaca gtctgatttc
9121 tccaaccgag ttgaaatttc caaagcaagt gcttctttaa atggggacct gagggcactc
9181 gctttgctcc tgtcagtaca cactcccaaa cagttaaacc cagctctaat tccaactctg
9241 caagagcttt taagcaaatg caggacttgt ctgcaacaga gaaactcact ccaagagcaa
9301 gaagccaaag aaagaaaaac taaagatgat gaaggagcaa ctcccattaa aaggcggcgt
9361 gttagcagtg atgaggagca cactgtagac agctgcatca gtgacatgaa aacagaaacc
9421 agggaggtcc tgaccccaac gagcacttct gacaatgaga ccagagactc ctcaattatt
9481 gatccaggaa ctgagcaaga tcttccttcc cctgaaaata gttctgttaa agaataccga
9541 atggaagttc catcttcgtt ttcagaagac atgtcaaata tcaggtcaca gcatgcagaa

9601 gaacagtcca acaatggtag atatgacgat tgtaaagaat ttaaagacct ccactgttcc
9661 aaggattcta ccctagctga ggaagaatct gagttccctt ctacttctat ctctgcagtt
9721 ctgtctgact tagctgactt gagaagctgt gatggccaag ctttgccctc ccaggaccct
9781 gaggttgctt tatctctcag ttgtggccat tccagaggac tctttagtca tatgcagcaa
9841 catgacattt tagataccct gtgtaggacc attgaatcta caatccatgt cgtcacaagg
9901 atatctggca aaggaaacca agctgcttct tga

Table 1(c)-USP N01 reference sequence (Derwent AAU82706)

Amino Acid sequence:

1 mcencadlve vlneisdveg gdglqlrkeh tlkiftyins wtqrqclccf keykhleifn
61 qvvcalinlv iaqvqvlrdq lckhcttini dstwqdesnq aeeplnidre cnegsterqk
121 siekksnstr icnlteeess kssdpfslws tdekeklllc vakifqiqfp lytaykhnth
181 ptiedistqe snilgafcdm ndvevplhll ryvclfcgkn glslmkdcfe ygtpetlpfl
241 iahafitvvs niriwlhipa vmqhiipfrt yvirylckls dqelrqsaar nmadlmwstv
301 kepldttlcf dkesldlafk yfmsptltmr laglsqitnq lhtfndvcnn eslvsdtets
361 iakeladwli snnvvehifg pnlhieiikq cqvilnflaa egrlstqhid ciwaaaqlkh
421 csryihdlfp slikn1dpvp lrhllnlvsa lepsvhteqt lylasmlika lwnnalaaka
481 qlskqssfas llntnipign kkeeeelrrt apspwspaas pqssdnsdth qsggsdiemd
541 eqlinrtkhv qqrlsdtees mqgssdetan sgedgssgpg sssghsdgss nevnsshasq
601 sagspgsevq sediadieal keededddhg hnppksscgt dlrnrklesq agiclgdsqg
661 tserngtssg tgkdlvfnte slpsvdnrmr mldacshsed pehdisgemn athiaqgsqe
721 scitrtgdfl getignelfn crqfigpqhh hhhhhhhhhh dghmvddmls addvscsssq
781 vsakseknma dfdgeesgce eelvqinsha eltshlqqhl pnlasiyheh lsqgpvvhkh
841 qfnsnavtdi nldnvckkgn tllwdivqde davnlsegli neaekllcsl vcwftdrqir
901 mrfiegclen lgnnrsvvis lrllpklfgt fqqfgssydt hwitmwaeke lnmmklffdn
961 lvyyiqtvre grqkhalysh saevqvrlqf ltcvfstlgs pdhfrlsleq vdilwhclve
1021 dsecyddalh wflnqvrskd qhamgmetyk hlflekmpql kpetismtgl nlfqhlcnla
1081 rlatsaydgc snselcgmdq fwgialraqs gdvsraaiqy insyyingkt glekeqefis
1141 kcmeslmias ssleqeshss lmviergllm lkthleafrr rfayhlrqwq iegtgisshl
1201 kalsdkqslp lrvvcqpagl pdkmtiemyp sdqvadlrae vthwyenlqk eqinqqaqlq
1261 efgqsnrkge fpgglmgpvr missgheltt dydekalhel gfkdmqmvfv slgaprrerk
1321 gegvqlpasc lpppqkdnip mllllqephl ttlfdllem1 asfkppsgkv avddseslrc
1381 eelhlhaenl srrvwellml lptcpnmlma fqnisdeqsf kaqsdhrsrh evshysmwll
1441 vswahccslv kssladsdhl qdwlkkltll ipetavrhes csglyklsls gldggdsinr
1501 sflllaastl lkflpdaqal kpiriddyee epilkpgcke yfwllcklvd nihikdasqt
1561 tlldldalar hladcirsre ildhqdgnve ddgltgllrl atsvvkhkpp fkfsregqef
1621 lrdifnllfl lpslkdrqqp kckshssraa aydllvemvk gsvenyrlih nwvmaqhmqs
1681 hapykwdywp hedvraecrf vgltnlgatc ylastiqqly mipearqavf takysedmkh
1741 kttllelqkm ftylmeseck aynprpfckt ytmdkqplnt geqkdmteff tdlitkieem
1801 spelkntvks lfggvitnnv vsldcehvsq taeefytvrc qvadmkniye sldevtikdt
1861 legdnmytcs qcgkkvraek racfkklpri xsfntmrytf nmvtmmkekv nthfsfplrl
1921 dmtpytedfl mgkserkegf kevsdhskds esyeydligv tvhtgtadgg hyysfirdiv
1981 nphayknnkw ylfndaevkp fdsaqlasec fggemttkty dsvtdkfmdf sfekthsaym
2041 lfykrmepee engreykfdv ssellewiwh dnmqflqdkn ifehtyfgfm wqlcscipst
2101 lpdpkavslm taklstsfvl etfihskekp tmlqwiellt kqfnnsqaac ewfldrmadd
2161 dwwpmqilik cpnqivrqmf qrlcihviqr lrpvhahlyl qpgmedgsdd mdtsvedigg
2221 rscvtrfvrt lllimehgvk phskhlteyf aflyefakmg eeesqfllsl qaistmvhfy
2281 mgtkgpenpq vevlseeegg eeeeeedils laeekyrpaa lekmialval lveqsrserh
2341 ltlsqtdmaa ltggkgfpfl fqhirdgini rqtcnlifsl crynnrlaeh ivsmlftsia
2401 kltpeaanpf fklltmlmef aggppgmppf asyilqriwe vieynpsqcl dwlavqtprn

```
2461 klahswvlqn menwverfll ahnyprvrts aayllvslip snsfrqmfrs trslhiptrd
2521 lplspdttvv lhqvynvllg llsraklyvd aavhgttklv pyfsfmtycl iskteklmfs
2581 tyfmdlwnlf qpklsepaia tnhnkqalls fwynvcadcp enirlivqnp vvtkniafny
2641 iladhddqdv vlfnrgmlpa yygilrlcce qspaftrqla shqniqwafk nltphasqyp
2701 gaveelfnlm qlfiaqrpdm reeeledikq fkkttiscyl rcldgrscwt tlisafrill
2761 esdedrllvv fnrglilmte sfntlhmmyh eatachvtgd lvellsifls vlkstrpylq
2821 rkdvkqaliq wqeriefahk lltllnsysp pelrnacidv lkelvllsph dflhtlvpfl
2881 qhnhctyhhs nipmslgpyf pcreniklig gksnirpprp elnmcllptm vetskgkddv
2941 ydrmlldyff syhqfihllc rvaincekft etlvklslvl ayeglplhla lfpklwtelc
3001 qtqsamsknc ikllcedpvf aeyikcilmd ertflnnniv ytfmthfllk vqsqvfsean
3061 canlistlit nlisqyqnlq sdfsnrveis kasaslngdl ralalllsvh tpkqlnpali
3121 ptlqellskc rtclqqrnsl qeqeakerkt kddegatpik rrrvssdeeh tvdscisdmk
3181 tetrevltpt stsdnetrds siidpgteqd lpspenssvk eyrmevpssf sedmsnirsq
3241 haeeqsnngr yddckefkdl hcskdstlae eesefpstsi savlsdladl rscdgqalps
3301 qdpevalsls cghsrglfsh mqqhdildtl crtiestihv vtrisgkgnq aas
```

Table 1(d)-USP N01 reference sequence (Derwent AAU82706)

Nucleotide sequence:

```
atgtgcgaga actgcgcaga cctggtggag gtgttaaatg aaatatcaga tgtagaaggt     60
ggtgatggac tgcagctcag aaaggaacat actctcaaaa tatttactta catcaattcc    120
tggacacaga ggcaatgtct atgctgcttc aaggaatata agcatttgga gatttttaat    180
caagtagtgt gtgcacttat taacttagtg attgcccaag ttcaagtgct ccgggaccag    240
ctttgtaaac attgtactac cattaacata gattccacgt ggcaagatga gagtaatcaa    300
gcagaagaac cactgaatat agatagagag tgtaatgaag gaagtacaga aagacaaaaa    360
tcaatagaaa aaaaatcaaa ctctacaaga atttgtaatc tgactgagga ggaatcttca    420
aagagttctg atcctttttag tttatggagt acagatgaga aggaaaaact cttactatgt    480
gtggcaaaaa tttttcaaat tcagtttccc ttatatactg cttacaagca taatactcac    540
cctactattg aggatatatc aactcaagaa agtaacatat taggggcatt ctgtgatatg    600
aatgatgtag aagtaccatt gcatttgctt cgttatgtat gtttgttttg tgggaaaaat    660
ggcctttctc tcatgaagga ttgctttgaa tatggaactc ctgaaacttt gccatttctt    720
atagcacatg cgtttattac agttgtgtct aatattagaa tatggctaca tattcccgct    780
gtcatgcagc acattatacc ttttaggacc tatgttatta ggtatttatg caagctctcg    840
gatcaggagt tacgacagag tgcagctcgt aacatggctg acttaatgtg gagcacagtc    900
aaagaaccat tggatacaac attatgcttt gataaagaaa gcctagatct tgcatttaag    960
tactttatgt cacctacttt gactatgagg ttggctggat tgagtcagat aacaaatcaa   1020
ctccatacct tcaatgatgt gtgcaataat gaatcattag tatcggacac agaaacgtcc   1080
attgcaaaag aacttgcaga ctggcttatt agcaacaatg tggtggagca tatatttgga   1140
ccaaatttac atattgagat tatcaaacag tgccaagtga ttttgaattt tttggcagca   1200
gaagggcgac tgagtactca acatattgac tgtatttggg ctgcagcaca gttgaaacat   1260
tgtagtcggt atatacatga cttatttcct tcactcatca agaatttgga tcccgtacca   1320
cttagacatc tacttaatct ggtctcagct cttgagccaa gtgttcatac tgaacagaca   1380
ctgtacttgg catccatgtt aattaaagca ctgtggaata acgcactagc agctaaggct   1440
cagttatcta aacagagttc ttttgcatct ttattaaata ctaatattcc cattggaaat   1500
aagaaagagg aagaagagct tagaagaaca gctccatcac cttggtcacc tgcagctagt   1560
cctcaaagca gtgataatag cgatacacat caaagtggag gtagtgacat tgaaatggat   1620
gagcaactta ttaatagaac caaacatgtg caacaacgac tttcagacac agaggaatcc   1680
atgcagggaa gttctgacga aactgccaac agtggtgaag atggaagcag tggtcctggt   1740
agcagtagtg ggcatagtga tggatctagc aatgaggtta attctagcca cgcaagccag   1800
tcagctggga gccctggcag tgaggtacag tcagaagaca ttgcagatat tgaagccctc   1860
aaagaggaag atgaagacga tgatcatggt cataatcctc ccaaaagcag ttgtggtaca   1920
gatcttcgga atagaaagtt agagagtcaa gcaggcattt gcctgggggga ctcccaaggc  1980
acgtcagaaa gaaatgggac aagcagcgga acaggaaagg acctggtttt taacactgaa   2040
tcattgccat cagtagataa tcgaatgcga atgctggatg cttgttcaca ctctgaagac   2100
ccagaacatg atatttcagg ggaaatgaat gctactcata tagcacaagg gtctcaggag   2160
tcttgtatca cacgaactgg ggacttcctt ggggagacta ttgggaatga attatttaat   2220
tgtcgacaat ttattggtcc acagcatcac caccaccacc accaccatca ccaccaccac   2280
gatgggcata tggttgatga tatgctaagt gcagatgatg tcagttgtag tagctcccag   2340
gttagtgcaa aatcagaaaa aaatatggct gattttgatg gtgaagaatc tggatgtgaa   2400
gaggagctag ttcagattaa ttcacatgcg gaactgacat ctcacctcca acaacatctt   2460
cccaatttag cttccattta ccatgaacat cttagtcaag gacctgtagt tcataaacat   2520
caattcaaca gtaatgctgt tacagacatt aatttggata atgtttgcaa gaaaggaaat   2580
actttgttgt gggatatagt ccaagatgaa gatgcagtta atctttctga aggattaata   2640
aatgaagcag agaaacttct ttgttcgtta gtatgttggt ttacagatag acaaattcga   2700
atgagattca ttgaaggttg ccttgaaaac ttgggaaaca acagatcagt agtaatttca   2760
cttcgtcttc ttccaaaact atttggtact tttcagcagt ttgggagcag ttacgataca   2820
cactggataa caatgtgggc agaaaaagaa ctgaacatga tgaagctttt ctttgataat   2880
ttggtatact acattcaaac tgtgagagaa ggaagacaaa aacatgcact gtacagccat   2940
```

```
agtgctgaag ttcaagttcg tcttcaattc ttgacttgtg tattttcaac tctgggatca     3000
cctgatcatt tcaggttaag tttagagcaa gttgacatct tatggcattg tttagtagaa     3060
gattctgaat gttatgatga tgcactccat tggtttttaa atcaagttcg aagtaaagat     3120
caacatgcta tgggtatgga aacctacaaa catcttttcc tggagaagat gccccagcta     3180
aaacctgaaa caattagcat gactggctta aacctgtttt cagcatctct gtaacttggc     3240
tcgattggct accagtgcct atgatggttg ttcaaattct gagctgtgtg gtatggacca     3300
attttggggc attgctttaa gagcacaatc tggtgatgtc agtcgagcag ctatccagta     3360
tattaactcc tattatatta atggtaaaac aggtttggag aaggagcaag aatttattag     3420
taagtgcatg gagagtctta tgatagcttc tagcagtctt gaacaggaat cacactcaag     3480
tctcatggtt atagaaagag gactccttat gctgaagaca catctggaag cgtttaggag     3540
aaggtttgca tatcatctga gacagtggca aattgaaggc actggtatta gtagtcattt     3600
gaaagcactg agtgacaaac agtctctgcc gctaagggtt gtatgccagc cagctggact     3660
tcctgacaag atgactattg aaatgtatcc tagtgaccag gtagcagatc ttagggctga     3720
agtaactcat tggtatgaaa atttacagaa agaacaaata aatcaacaag ctcagcttca     3780
ggagtttggt caaagcaacc gaaaaggaga gtttcctgga ggcctcatgg gacctgtcag     3840
gatgatttca tctggacacg agttaacaac agattatgat gaaaaagcac ttcatgagct     3900
tggttttaag gatatgcaga tggtatttgt atctttgggt gcaccaagga gagagcggaa     3960
aggggaaggt gttcagctgc cagcatcttg cctcccaccc cctcagaagg acaacattcc     4020
aatgcttttg cttttacaag agcctcattt aactactctt tttgatttat tagagatgct     4080
tgcatcattt aaaccaccct caggaaaagt ggcagtggat gatagtgaga gcttacgatg     4140
tgaagaactt catcttcatg cagaaaatct gtctaggcgg gtctgggagc tactgatgct     4200
tcttcctaca tgtcctaata tgttgatggc attccagaat atctcagatg agcagagttt     4260
taaagctcag tctgatcaca ggtctagaca tgaagtttca cattattcaa tgtggctctt     4320
ggtgagttgg gctcattgct gttctttagt gaaatctagc cttgctgata gcgatcattt     4380
acaagattgg ctaaagaaat tgactctcct tattcctgag actgcagttc gtcatgaatc     4440
atgcagtggt ctctataagt tatccctgtc agggctggat ggaggagact caatcaatcg     4500
ttcttttctg ctattggctg cctcaacatt attgaaattt cttcctgatg ctcaagcact     4560
caaacctatt aggatagatg attatgagga agaaccaata ttaaaaccag gatgtaaaga     4620
gtatttttgg ttgttatgca aattagttga caacatacat ataaaggacg ctagtcagac     4680
aacgctcctc gacttagatg ccttggcaag acatttggct gactgtattc gaagtaggga     4740
gatccttgat catcaggatg gtaatgtaga agatgatggg cttacaggac tcctaaggct     4800
tgcaacaagt gttgttaaac acaaaccacc ctttaaattt tcaagggaag gacaggaatt     4860
tttgagagat atcttcaatc tcctgttttt gttgccaagt ctaaaggacc gacaacagcc     4920
aaagtgcaaa tcacattctt caagagctgc cgcttacgat ttgttagtag agatggtaaa     4980
ggggtctgtt gagaactaca ggctaataca caactgggtt atggcacaac acatgcagtc     5040
ccatgcacct tataaatggg attactggcc tcatgaagat gtccgtgctg aatgtagatt     5100
tgttggcctt actaaccttg gagctacttg ttacttagct tctactattc agcaacttta     5160
tatgatacct gaggcaagac aggctgtctt cactgccaag tattcagagg atatgaagca     5220
caagaccact cttctggagc ttcagaaaat gtttacatat ttaatggaga gtgaatgcaa     5280
agcatataat cctagacctt tctgtaaaac atacaccatg gataagcagc ctctgaatac     5340
tggggaacag aaagatatga cagagttttt tactgatcta attaccaaaa tcgaagaaat     5400
gtctcccgaa ctgaaaaata ccgtcaaaag tttatttgga ggtgtaatta caaacaatgt     5460
tgtatccttg gattgtgaac atgttagtca aactgctgaa gagttttata ctgtgaggtg     5520
ccaagtggct gatatgaaga acatttatga atctcttgat gaagttacta taaaagacac     5580
tttggaaggt gataacatgt atacttgttc tcaatgtggg aagaaagtac gagctgaaaa     5640
aagggcatgt tttaagaaat tgcctcgcat tttnagtttc aatactatga gatacacatt     5700
taatatggtc acgatgatga aagagaaagt gaatacacac ttttccttcc cattacgttt     5760
ggacatgacg ccctatacag aagatttttct tatgggaaag agtgagagga aagaaggttt     5820
taaagaagtc agtgatcatt caaaagactc agagagctat gaatatgact tgataggagt     5880
gactgttcac acaggaacgg cagatggtgg acactattat agctttatca gagatatagt     5940
aaatccccat gcttataaaa acaataaatg gtatcttttt aatgatgctg aggtaaaacc     6000
ttttgattct gctcaacttg catctgaatg ttttggtgga gagatgacga ccaagaccta     6060
tgattctgtt acagataaat ttatggactt ctcttttgaa aagacacaca gtgcatatat     6120
gctgttttac aaacgcatgg aaccagagga agaaaatggc agagaataca aatttgatgt     6180
ttcgtcagag ttactagagt ggatttggca tgataacatg cagtttcttc aagacaaaaa     6240
catttttgaa catacatatt ttggatttat gtggcaattg tgtagttgta ttcccagtac     6300
attaccagat cctaaagctg tgtccttaat gacagcaaag ttaagcactt cctttgtcct     6360
agagacattt attcattcta aagaaaagcc cacgatgctt cagtggattg aactgttgac     6420
gaaacagttt aataatagtc aggcagcttg tgagtggttt ttagatcgta tggctgatga     6480
cgactggtgg ccaatgcaga tactaattaa gtgccctaat caaattgtga gacagatgtt     6540
tcagcgtttg tgtatccatg tgattcagag gctgagacct gtgcatgctc atctctattt     6600
gcagccagga atggaagatg ggtcagatga tatggatacc tcagtagaag atattggtgg     6660
tcgttcatgt gtcactcgct ttgtgagaac cctgttatta attatggaac atggtgtaaa     6720
acctcacagt aaacatctta cagagtattt tgccttcctt tacgaatttg caaaaatggg     6780
tgaagaagag agccaatttt tgctttcatt gcaagctata tctacaatgg tacattttta     6840
catgggaaca aaaggacctg aaaatcctca agttgaagtg ttatcagagg aagaaggggg     6900
agaagaagag gaggaagaag atatcctctc tctggcagaa gaaaaataca ggccagctgc     6960
ccttgaaaag atgatagctt tagttgctct tttggttgaa cagtctcgat cagaaaggca     7020
tttgacatta tcacagactg acatggcagc attaacagga ggaaagggat ttcccttctt     7080
gtttcaacat attcgtgatg gcatcaatat aagacaaact tgtaatctga ttttcagcct     7140
gtgtcgatac aataatcgac ttgcagaaca tattgtatct atgcttttca catcaatagc     7200
aaagttgact cctgaggcag ccaatccttt ctttaagttg ttgactatgc taatggagtt     7260
tgctggtgga cctccaggaa tgcctcccct tgcatcttat attctgcaga ggatatggga     7320
ggtgattgaa tacaatcctt ctcagtgtct agattggttg gcagtgcaga caccccgaaa     7380
taaactggca cacagctggg tcttacagaa tatggaaaac tgggtcgagc ggtttctttt     7440
ggctcacaat tatcctagag tgaggacttc tgcagcttat cttctggtgt cccttatacc     7500
```

```
aagcaattca ttccgtcaga tgttccggtc aacaaggtct ttgcacatcc caacccgtga     7560
ccttccactc agtccagaca caacagtagt cctacatcag gtctacaacg tgctccttgg     7620
tttgctctca agagccaaac tttatgttga tgctgctgtt catggcacta caaagctagt     7680
gccctatttt agctttatga cttactgttt aatttccaaa actgagaagc tgatgttttc     7740
cacatatttc atggatttgt ggaacctttt ccagcctaaa ctttctgagc cagcaatagc     7800
tacaaatcac aataaacagg ctttgctttc attttggtac aatgtctgtg ctgactgtcc     7860
agagaatatc cgccttattg ttcagaaccc agtggtaacc aagaacattg ccttcaatta     7920
catccttgct gaccatgatg atcaggatgt ggtgcttttt aaccgtggga tgctgccagc     7980
gtactatggc attctgaggc tctgctgtga gcagtctcct gcattcacac gacaactggc     8040
ttctcaccag aacatccagt gggcctttaa gaatcttaca ccacatgcca gccaataccc     8100
tggagcagta gaagaactgt ttaacctgat gcagctgttt atagctcaga ggccagatat     8160
gagagaagaa gaattagaag atattaaaca gttcaagaaa acaaccataa gttgttactt     8220
acgttgctta gatggccgct cctgctggac tactttaata agtgccttca gaatactatt     8280
agaatctgat gaagacagac ttcttgttgt atttaatcga ggattgattc taatgacaga     8340
gtctttcaac actttgcaca tgatgtatca cgaagctaca gcttgccatg tgactggaga     8400
tttagtagaa cttctgtcaa tatttctttc ggttttgaag tctacacgcc cttatcttca     8460
gagaaaagat gtgaaacaag cattaatcca gtggcaggag cgaattgaat ttgcccataa     8520
actgttaact cttcttaatt cctatagtcc tccagaactt agaaatgcct gtatagatgt     8580
cctcaaggaa cttgtacttt tgagtcccca tgattttctt catactctgg ttccctttct     8640
acaacacaac cattgtactt accatcacag taatatacca atgtctcttg gaccttattt     8700
cccttgtcga gaaaatatca agctaatagg agggaaaagc aatattcggc ctccgcgccc     8760
tgaactcaat atgtgcctct tgcccacaat ggtggaaacc agtaagggca aagatgacgt     8820
ttatgatcgt atgctgctag actacttctt ttcttatcat cagttcatcc atctattatg     8880
ccgagttgca atcaactgtg aaaaatttac tgaaacatta gttaagctga gtgtcctagt     8940
tgcctatgaa ggtttgccac ttcatcttgc actgttcccc aaactttgga ctgagctatg     9000
ccagactcag tctgctatgt caaaaaactg catcaagctt ttgtgtgaag atcctgtttt     9060
cgcagaatat attaaatgta tcctaatgga tgaaagaact tttttaaaca acaacattgt     9120
ctacacgttc atgacacatt tccttctaaa ggttcaaagt caagtgtttt ctgaagcaaa     9180
ctgtgccaat ttgatcagca ctcttattac aaacttgata agccagtatc agaacctaca     9240
gtctgatttc tccaaccgag ttgaaatttc caaagcaagt gcttctttaa atggggacct     9300
gagggcactc gctttgctcc tgtcagtaca cactcccaaa cagttaaacc cagctctaat     9360
tccaactctg caagagcttt taagcaaatg caggacttgt ctgcaacaga gaaactcact     9420
ccaagagcaa gaagccaaag aaagaaaaac taaagatgat gaaggagcaa ctcccattaa     9480
aaggcggcgt gttagcagtg atgaggagca cactgtagac agctgcatca gtgacatgaa     9540
aacagaaacc agggaggtcc tgaccccaac gagcacttct gacaatgaga ccagagactc     9600
ctcaattatt gatccaggaa ctgagcaaga tcttccttcc cctgaaaata gttctgttaa     9660
agaataccga atggaagttc catcttcgtt ttcagaagac atgtcaaata tcaggtcaca     9720
gcatgcagaa gaacagtcca acaatggtag atatgacgat tgtaaagaat ttaaagacct     9780
ccactgttcc aaggattcta ccctagctga ggaagaatct gagttccctt ctacttctat     9840
ctctgcagtt ctgtctgact tagctgactt gagaagctgt gatggccaag ctttgccctc     9900
ccaggaccct gaggttgctt tatctctcag ttgtggccat tccagaggac tctttagtca     9960
tatgcagcaa catgacattt tagataccct gtgtaggacc attgaatcta caatccatgt    10020
cgtcacaagg atatctggca aaggaaacca agctgcttct tga                      10063
```

Table 2(a)USP_N07: novel splice variant- amino acid sequence

```
LOCUS           USP_N07      1355 aa     PRT
Accession    NP_060414
GeneSeq         AAU82714
ORIGIN         human

Splice form characterized by 1 insertions:
Pos 14 - Pos 81
   1 MVPGEENQLV PKEIENAAEE PRVLCIIQDT TNSKTVNERI TLNLPASTPV RKLFEDVANK
  61 VGYINGTFDL VWGNGINTAD MAPLDHTSDK SLLDANFEPG KKNFLHLTDK DGEQPQILLE
 121 DSSAGEDSVH DRFIGPLPRE GSVGSTSDYV SQSYSYSSIL NKSETGYVGL VNQAMTCYLN
 181 SLLQTLFMTP EFRNALYKWE FEESEEDPVT SIPYQLQRLF VLLQTSKKRA IETTDVTRSF
 241 GWDSSEAWQQ HDVQELCRVM FDALEQKWKQ TEQADLINEL YQGKLKDYVR CLECGYEGWR
 301 IDTYLDIPLV IRPYGSSQAF ASVEEALHAF IQPEILDGPN QYFCERCKKK CDARKGLRFL
 361 HFPYLLTLQL KRFDFDYTTM HRIKLNDRMT FPEELDMSTF IDVEDEKSPQ TESCTDSGAE
 421 NEGSCHSDQM SNDFSNDDGV DEGICLETNS GTEKISKSGL EKNSLIYELF SVMVHSGSAA
 481 GGHYYACIKS FSDEQWYSFN DQHVSRITQE DIKKTHGGSS GSRGYYSSAF ASSTNAYMLI
 541 YRLKDPARNA KFLEVDEYPE HIKNLVQKER ELEEQEKRQR EIERNTCKIK LFCLHPTKQV
 601 MMENKLEVHK DKTLKEAVEM AYKMMDLEEV IPLDCCRLVK YDEFHDYLER SYEGEEDTPM
 661 GLLLGGVKST YMFDLLLETR KPDQVFQSYK PGEVMVKVHV VDLKAESVAA PITVRAYLNQ
 721 TVTEFKQLIS KAIHLPAETM RIVLERCYND LRLLSVSSKT LKAEGFFRSN KVFVESSETL
 781 DYQMAFADSH LWKLLDRHAN TIRLFVLLPE QSPVSYSKRT AYQKAGGDSG NVDDDCERVK
 841 GPVGSLKSVE AILEESTEKL KSLSLQQQQD GDNGDSSKST ETSDFENIES PLNERDSSAS
 901 VDNRELEQHI QTSDPENFQS EERSDSDVNN DRSTSSVDSD ILSSSHSSDT LCNADNAQIP
 961 LANGLDSHSI TSSRRTKANE GKKETWDTAE EDSGTDSEYD ESGKSRGEMQ YMYFKAEPYA
```

```
1021 ADEGSGEGHK WLMVHVDKRI TLAAFKQHLE PFVGVLSSHF KVFRVYASNQ EFESVRLNET
1081 LSSFSDDNKI TIRLGRALKK GEYRVKVYQL LVNEQEPCKF LLDAVFAKGM TVRQSKEELI
1141 PQLREQCGLE LSIDRFRLRK KTWKNPGTVF LDYHIYEEDI NISSNWEVFL EVLDGVEKMK
1201 SMSQLAVLSR RWKPSEMKLD PFQEVVLESS SVDELREKLS EISGIPLDDI EFAKGRGTFP
1261 CDISVLDIHQ DLDWNPKVST LNVWPLYICD DGAVIFYRDK TEELMELTDE QRNELMKKES
1321 SRLQKTGHRV TYSPRKEKAL KIYLDGAPNK DLTQD
```

Table 2(b)USP N07: novel splice variant- nucleotide sequence

```
   1 atggtgcccg gcgaggagaa ccaactggtc ccgaaagaga tagaaaatgc tgctgaagaa
  61 cctagagtct tatgtattat acaagatact actaattcaa agacagtgaa tgaacggatc
 121 actttaaatt taccagcatc tactccagtc agaaagctct ttgaagatgt ggccaacaaa
 181 gtaggctaca taaatggaac ctttgacttg gtgtggggaa atggaatcaa tactgctgat
 241 atggcaccac tggatcatac cagtgacaag tcacttctcg acgctaattt tgagccagga
 301 aagaagaact ttctgcattt gacagataaa gatggtgaac aacctcaaat actgctggag
 361 gattccagtg ctggggaaga cagtgttcat gacaggttta taggtccgct tccaagagaa
 421 ggttctgtgg gttctaccag tgattatgtc agccaaagct actcctactc atctattttg
 481 aataaatcag aaactggata tgtgggacta gtaaaccaag caatgacttg ctatttgaat
 541 agccttttgc aaacactttt tatgactcct gaatttagga atgcattata taagtgggaa
 601 tttgaagaat ctgaagaaga tccagtgaca agtattccat accaacttca aaggcttttt
 661 gttttgttac aaaccagcaa aaagagagca attgaaacca cagatgttac aaggagcttt
 721 ggatgggata gtagtgaggc ttggcagcag catgatgtac aagaactatg cagagtcatg
 781 tttgatgctt tggaacagaa atggaagcaa acagaacagg ctgatcttat aaatgagcta
 841 tatcaaggca agctgaagga ctacgtgaga tgtctggaat gtggttatga gggctggcga
 901 atcgacacat atcttgatat tccattggtc atccgacctt atgggtccag ccaagcattt
 961 gctagtgtgg aagaagcatt gcatgcattt attcagccag agattctgga tggcccaaat
1021 cagtattttt gtgaacgttg taagaagaag tgtgatgcac ggaagggcct tcggtttttg
1081 cattttcctt atctgctgac cttacagctg aaaagattcg atttgatta tacaaccatg
1141 cataggatta aactgaatga tcgaatgaca tttcccgagg aactagatat gagtactttt
1201 attgatgttg aagatgagaa atctcctcag actgaaagtt gcactgacag tggagcagaa
1261 aatgaaggta gttgtcacag tgatcagatg agcaacgatt tctccaatga tgatggtgtt
1321 gatgaaggaa tctgtcttga aaccaatagt ggaactgaaa agatctcaaa atctggactt
1381 gaaaagaatt ccttgatcta tgaacttttc tctgttatgg ttcattctgg gagcgctgct
1441 ggtggtcatt attatgcatg tataaagtca ttcagtgatg agcagtggta cagcttcaat
1501 gatcaacatg tcagcaggat aacacaagag gacattaaga aaacacatgg tggatcttca
1561 ggaagcagag gatattattc tagtgctttc gcaagttcca caaatgcata tatgctgatc
1621 tatagactga aggatccagc cagaaatgca aaatttctag aagtggatga atacccagaa
1681 catattaaaa acttggtgca gaaagagaga gagttggaag aacaagaaaa gagacaacga
1741 gaaattgagc gcaatacatg caagataaaa ttattctgtt tgcatcctac aaaacaagta
1801 atgatggaaa ataaattgga ggttcataag gataagacat taaaggaagc agtagaaatg
1861 gcttataaga tgatggattt agaagaggta atacccctgg attgctgtcg ccttgttaaa
1921 tatgatgagt ttcatgatta tctagaacgg tcatatgaag gagaagaaga tacaccaatg
1981 gggcttctac taggtggcgt caagtcaaca tatatgtttg atctgctgtt ggagacgaga
2041 aagcctgatc aggttttcca atcttataaa cctggagaag tgatggtgaa agttcatgtt
2101 gttgatctaa aggcagaatc tgtagctgct cctataactg ttcgtgctta cttaaatcag
2161 acagttacag aattcaaaca actgatttca aaggccatcc atttacctgc tgaaacaatg
2221 agaatagtgc tggaacgctg ctacaatgat ttgcgtcttc tcagtgtctc cagtaaaacc
2281 ctgaaagctg aaggattttt tagaagtaac aaggtgtttg ttgaaagctc cgagactttg
2341 gattaccaga tggcctttgc agactctcat ttatggaaac tcctggatcg gcatgcaaat
2401 acaatcagat tatttgtttt gctacctgaa caatccccag tatcttattc caaaaggaca
2461 gcataccaga aagctggagg cgattctggt aatgtggatg atgactgtga aagagtcaaa
2521 ggacctgtag gaagcctaaa gtctgtggaa gctattctag aagaaagcac tgaaaaactc
2581 aaaagcttgt cactgcagca acagcaggat ggagataatg gggacagcag caaaagtact
2641 gagacaagtg actttgaaaa catcgaatca cctctcaatg agagggactc ttcagcatca
2701 gtggataata gagaacttga acagcatatt cagacttctg atccagaaaa ttttcagtct
```

```
2761 gaagaacgat cagactcaga tgtgaataat gacaggagta caagttcagt ggacagtgat
2821 attcttagct ccagtcatag cagtgatact ttgtgcaatg cagacaatgc tcagatccct
2881 ttggctaatg gacttgactc tcacagtatc acaagtagta gaagaacgaa agcaaatgaa
2941 gggaaaaaag aaacatggga tacagcagaa gaagactctg gaactgatag tgaatatgat
3001 gagagtggca agagtagggg agaaatgcag tacatgtatt tcaaagctga accttatgct
3061 gcagatgaag gttctgggga aggacataaa tggttgatgg tgcatgttga taaaagaatt
3121 actctggcag ctttcaaaca acatttagag ccctttgttg gagttttgtc ctctcacttc
3181 aaggtctttc gagtgtatgc cagcaatcaa gagtttgaga gcgtccggct gaatgagaca
3241 ctttcatcat ttctgatga caataagatt acaattagac tggggagagc acttaaaaaa
3301 ggagaataca gagttaaagt ataccagctt ttggtcaatg aacaagagcc atgcaagttt
3361 ctgctagatg ctgtgtttgc taaaggaatg actgtacggc aatcaaaaga ggaattaatt
3421 cctcagctca gggagcaatg tggtttagag ctcagtattg acaggtttcg tctaaggaaa
3481 aaaacatgga agaatcctgg cactgtcttt ttggattatc atatttatga agaagatatt
3541 aatatttcca gcaactggga ggttttcctt gaagttcttg atggggtaga gaagatgaag
3601 tccatgtcac agcttgcagt tttgtcaaga cggtggaagc cttcagagat gaagttggat
3661 cccttccagg aggttgtatt ggaaagcagt agtgtggacg aattgcgaga gaagcttagt
3721 gaaatcagtg ggattccttt ggatgatatt gaatttgcta agggtagagg aacatttccc
3781 tgtgatattt ctgtccttga tattcatcaa gatttagact ggaatcctaa agtttctacc
3841 ctgaatgtct ggcctcttta tatctgtgat gatggtgcgg tcatatttta tagggataaa
3901 acagaagaat taatggaatt gacagatgag caaagaaatg aactgatgaa aaaagaaagc
3961 agtcgactcc agaagactgg acatcgtgta acatactcac ctcgtaaaga gaaagcacta
4021 aaaatatatc tggatggagc accaaataaa gatctgactc aagactga
```

Table 2(c)-USP N07 reference sequence (Derwent AAU82714)

Amino Acid sequence:

```
   1 mvpgeenqlv pkeapldhts dkslldanfe pgkknflhlt dkdgeqpqil ledssageds
  61 vhdrfigplp regsvgstsd yvsqsysyss ilnksetgyv glvnqamtcy lnsllqtlfm
 121 tpefrnalyk wefeeseedp vtsipyqlqr lfvllqtskk raiettdvtr sfgwdsseaw
 181 qqhdvqelcr vmfdaleqkw kqteqadlin elyqgklkdy vrclecgyeg wridtyldip
 241 lvirpygssq afasveealh afiqpeildg pnqyfcerck kkcdarkglr flhfpylltl
 301 qlkrfdfdyt tmhriklndr mtfpeeldms tfidvedeks pqtesctdsg aenegschsd
 361 qmsndfsndd gvdegiclet nsgtekisks gleknsliye lfsvmahsgs aagghyyaci
 421 ksfsdeqwys fddqhvsrit qedikkthgg ssgsrgyyss afasstnaym liyrlkdpar
 481 nakflevgey pehiknlvqk ereleeqekr qreierntck iklfclhptk qvmmenklev
 541 hkdktlkeav emaykmmdle evipldccrl vkydefhdyl ersyegeedt pmglllggvk
 601 stymfdllle trkpdqvfqs ykpgevmvkv hvvdlkaesv aapitvrayl nqtvtefkql
 661 iskaihlpae tmrivlercy ndlrllsvss ktlkaegffr snkvfvesse tldyqmafad
 721 shlwklldrh antirlfvll peqspvsysk rtayqkaggd sgnvdddcer vkgpvgslks
 781 veaileeste klkslslqqq qdgdngdssk stetsdfeni esplnerdss asvdnreleq
 841 hiqtsdpenf qseersdsdv nndrstssvd sdilssshss dtlcnadnaq iplangldsh
 901 sitssrrtka negkketwdt aeedsgtdse ydesgksrge mqymyfkaep yaadegsgeg
 961 hkwlmvhvdk ritlaafkqh lepfvgvlss hfkvfrvyas nqefesvrln etlssfsddn
1021 kitirlgral kkgeyrvkvy qllvneqepc kflldavfak gmtvrqskee lipqlreqcg
1081 lelsidrfrl rkktwknpgt vfldyhiyee dinissnwev flevldgvek mksmsqlavl
1141 srrwkpsemk ldpfqevvle sssvdelrek lseisgipld diefakgrgt fpcdisvldi
1201 hqdldwnpkv stlnvwplyi cddgavifyr dkteelmelt deqrnelmkk essrlqktgh
1261 rvtysprkek alkiyldgap nkdltqd
```

Table 2(d)-USP N07 reference sequence (Derwent AAU82714)

Nucleotide sequence:

```
atggtgcccg gcgaggagaa ccaactggtc ccgaaagagg caccactgga tcataccagt   60
gacaagtcac ttctcgacgc taattttgag ccaggaaaga agaactttct gcatttgaca  120
```

```
gataaagatg gtgaacaacc tcaaatactg ctggaggatt ccagtgctgg ggaagacagt    180
gttcatgaca ggtttatagg tccgcttcca agagaaggtt ctgtgggttc taccagtgat    240
tatgtcagcc aaagctactc ctactcatct attttgaata aatcagaaac tggatatgtg    300
ggactagtaa accaagcaat gacttgctat ttgaatagcc ttttgcaaac actttttatg    360
actcctgaat ttaggaatgc attatataag tgggaatttg aagaatctga agaagatcca    420
gtgacaagta ttccatacca acttcaaagg ctttttgttt tgttacaaac cagcaaaag     480
agagcaattg aaaccacaga tgttacaagg agctttggat gggatagtag tgaggcttgg    540
cagcagcatg atgtacaaga actatgcaga gtcatgtttg atgctttgga acagaaatgg    600
aagcaaacag aacaggctga tcttataaat gagctatatc aaggcaagct gaaggactac    660
gtgagatgtc tggaatgtgg ttatgagggc tggcgaatcg acacatatct tgatatccca    720
ttggtcatcc gaccttatgg gtccagccaa gcatttgcta gtgtggaaga agcattgcat    780
gcatttattc agccagagat tctggatggc ccaaatcagt atttttgtga acgttgtaag    840
aagaagtgtg atgcacggaa gggccttcgg tttttgcatt ttccttatct gctgacctta    900
cagctgaaaa gattcgattt tgattataca accatgcata ggattaaact gaatgatcga    960
atgacatttc ccgaggaact agatatgagt actttattg atgttgaaga tgagaaatct    1020
cctcagactg aaagttgcac tgacagtgga gcagaaaatg aaggtagttg tcacagtgat   1080
cagatgagca acgatttctc caatgatgat ggtgttgatg aaggaatctg tcttgaaacc   1140
aatagtggaa ctgaaaagat ctcaaaatct ggacttgaaa agaattcctt gatctatgaa   1200
ctttctctg ttatggctca ttctgggagc gctgctggtg gtcattatta tgcatgtata    1260
aagtcattca gtgatgagca gtggtacagc ttcgatgatc aacatgtcag caggataaca   1320
caagaggaca ttaagaaaac acatggtgga tcttcaggaa gcagaggata ttattctagt   1380
gctttcgcaa gttccacaaa tgcatatatg ctgatctata gactgaagga tccagccaga   1440
aatgcaaaat ttctagaagt gggtgaatac ccagaacata ttaaaaactt ggtgcagaaa   1500
gagagagagt tggaagaaca agaaaagaga caacgagaaa ttgagcgcaa tacatgcaag   1560
ataaaattat tctgtttgca tcctacaaaa caagtaatga tggaaaataa attggaggtt   1620
cataaggata agacattaaa ggaagcagta gaaatggctt ataagatgat ggatttagaa   1680
gaggtaatac ccctggattg ctgtcgcctt gttaaatatg atgagtttca tgattatcta   1740
gaacggtcat atgaaggaga agaagataca ccaatggggc ttctactagg tggcgtcaag   1800
tcaacatata tgtttgatct gctgttggag acgagaaagc ctgatcaggt tttccaatct   1860
tataaacctg gagaagtgat ggtgaaagtt catgttgttg atctaaaggc agaatctgta   1920
gctgctccta taactgttcg tgcttactta aatcagacag ttacagaatt caaacaactg   1980
atttcaaagg ccatccattt acctgctgaa acaatgagaa tagtgctgga acgctgctac   2040
aatgatttgc gtcttctcag tgtctccagt aaaaccctga aagctgaagg atttttttaga  2100
agtaacaagg tgtttgttga aagctccgag actttggatt accagatggc ctttgcagac   2160
tctcatttat ggaaactcct ggatcggcat gcaaatacaa tcagattatt tgttttgcta   2220
cctgaacaat ccccagtatc ttattccaaa aggacagcat accagaaagc tggaggcgat   2280
tctggtaatg tggatgatga ctgtgaaaga gtcaaaggac ctgtaggaag cctaaagtct   2340
gtggaagcta ttctagaaga aagcactgaa aaactcaaaa gcttgtcact gcagcaacag   2400
caggatggag ataatgggga cagcagcaaa agtactgaga caagtgactt tgaaaacatc   2460
gaatcacctc tcaatgagag ggactcttca gcatcagtgg ataatagaga acttgaacag   2520
catattcaga cttctgatcc agaaaatttt cagtctgaag aacgatcaga ctcagatgtg   2580
aataatgaca ggagtacaag ttcagtggac agtgatattc ttagctccag tcatagcagt   2640
gatactttgt gcaatgcaga caatgctcag atccctttgg ctaatggact tgactctcac   2700
agtatcacaa gtagtagaag aacgaaagca aatgaaggga aaaaagaaac atgggataca   2760
gcagaagaag actctggaac tgatagtgaa tatgatgaga gtggcaagag taggggagaa   2820
atgcagtaca tgtatttcaa agctgaacct tatgctgcag atgaaggttc tggggaagga   2880
cataaatggt tgatggtgca tgttgataaa agaattactc tggcagcttt caaacaacat   2940
ttagagccct ttgttggagt tttgtcctct cacttcaagg tctttcgagt gtatgccagc   3000
aatcaagagt ttgagagcgt ccggctgaat gagacacttt catcattttc tgatgacaat   3060
aagattacaa ttagactggg gagagcactt aaaaaaggag aatacagagt taaagtatac   3120
cagcttttgg tcaatgaaca agagccatgc aagtttctgc tagatgctgt gtttgctaaa   3180
ggaatgactg tacggcaatc aaaagaggaa ttaattcctc agctcaggga gcaatgtggt   3240
ttagagctca gtattgacag gtttcgtcta aggaaaaaaa catggaagaa tcctggcact   3300
gtcttttttgg attatcatat ttatgaagaa gatattaata tttccagcaa ctgggaggtt  3360
ttccttgaag ttcttgatgg ggtagagaag atgaagtcca tgtcacagct tgcagtttg    3420
tcaagacggt ggaagccttc agagatgaag ttggatccct tccaggaggt tgtattggaa   3480
agcagtagtg tggacgaatt gcgagagaag cttagtgaaa tcagtgggat tcctttggat   3540
gatattgaat ttgctaaggg tagaggaaca tttccctgtg atatttctgt ccttgatatt   3600
catcaggatt tagactggaa tcctaaagtt tctaccctga atgtctggcc tctttatatc   3660
tgtgatgatg gtgcggtcat attttatagg gataaaacag aagaattaat ggaattgaca   3720
gatgagcaaa gaaatgaact gatgaaaaaa gaaagcagtc gactccagaa gactggacat   3780
cgtgtaacat actcacctcg taaagagaaa gcactaaaaa tatatctgga tggagcacca   3840
aataaagatc tgactcaaga ctga                                           3864
```

Table 3(a)-USP N11: novel splice variant: amino acid sequence

```
LOCUS           USP_N11        402 aa      PRT
Accession    AK022614
GeneSeq         AAU82713
ORIGIN        human

Splice form characterized by 1 insertions:
Pos 12 - Pos 48
        1 MTVRNIASIC NMEEPPALGS PGWTLLAPPL VRAFGELRLE EGIAVPCRGT NASALEKDIG
```

```
 61 PEQFPINEHY FGLVNFGNTC YCNSVLQALY FCRPFRENVL AYKAQQKKKE NLLTCLADLF
121 HSIATQKKKV GVIPPKKFIS RLRKENDLFD NYMQQDAHEF LNYLLNTIAD ILQEEKKQEK
181 QNGKLKNGNM NEPAENNKPE LTWVHEIFQG TLTNETRCLN CETVSSKDED FLDLSVDVEQ
241 NTSITHCLRD FSNTETLCSE QKYYCETCCS KQEAQKRMRV KKLPMILALH LKRFKYMEQL
301 HRYTKLSYRV VFPLELRLFN TSSDAVNLDR MYDLVAVVVH CGSGPNRGHY ITIVKSHGFW
361 LLFDDDIVEK IDAQAIEEFY GLTSDISKNS ESGYILFYQS RE
```

Table 3(b)- USP N11: novel splice variant: nucleotide sequence

```
   1 atgactgtcc gaaacatcgc ctccatctgt aatatgggca ccaatgcctc tgctctggaa
  61 aaagacattg gtccagagca gtttccaatc aatgaacact atttcggatt ggtcaatttt
 121 ggaaacacat gctactgtaa ctccgtgctt caggcattgt acttctgccg tccattccgg
 181 gagaatgtgt tggcatacaa ggcccagcaa aagaagaagg aaaacttgct gacgtgcctg
 241 gcggaccttt tccacagcat tgccacacag aagaagaagg ttggcgtcat cccaccaaag
 301 aagttcattt caaggctgag aaaagagaat gatctctttg ataactacat gcagcaggat
 361 gctcatgaat ttttaaatta tttgctaaac actattgcgg acatccttca ggaggagaag
 421 aaacaggaaa aacaaaatgg aaaattaaaa aatggcaaca tgaacgaacc tgcggaaaat
 481 aataaaccag aactcacctg ggtccatgag atttttcagg gaacgcttac caatgaaact
 541 cgatgcttga actgtgaaac tgttagtagc aaagatgaag attttcttga cctttctgtt
 601 gatgtggagc agaatacatc cattacccac tgtctaagag acttcagcaa cacagaaaca
 661 ctgtgtagtg aacaaaaata ttattgtgaa acatgctgca gcaaacaaga agcccagaaa
 721 aggatgaggg taaaaaagct gcccatgatc ttggccctgc acctaaagcg gttcaagtac
 781 atggagcagc tgcacagata caccaagctg tcttaccgtg tggtcttccc tctggaactc
 841 cggctcttca acacctccag tgatgcagtg aacctggacc gcatgtatga cttggttgcg
 901 gtggtcgttc actgtggcag tggtcctaat cgtgggcatt atatcactat tgtgaaaagt
 961 cacggcttct ggcttttgtt tgatgatgac attgtagaga aaatagatgc tcaagctatt
1021 gaagaattct atggcctgac gtcagatata tcaaaaaatt cagaatctgg atatatttta
1081 ttctatcagt caagagagta a
```

Table 3(c)-USP N11 reference sequence (Derwent AAU82713)

Amino Acid sequence:

```
  1 mtvrniasic nmgtnasale kdigpeqfpi nehyfglvnf gntcycnsvl qalyfcrpfr
 61 envlaykaqq kkkenlltcl adlfhsiatq kkkvgvippk kfisrlrken dlfdnymqqd
121 aheflnylln tiadilqeek kqekqngklk ngnmnepaen nkpeltwvhe ifqgtltnet
181 rclncetvss kdedfldlsv dveqntsith clrdfsntet lcseqkyyce tccskqeaqk
241 rmrvkklpmv lalhlkrfky meqlrrytkl syrvvfplel rlfntssdav nldrmydlva
301 vvvhcgsgpn rghyitivks hgfwllfddd ivekidaqai eefygltsdi sknsesgyil
361 fyqsre
```

Table 3(d)-USP N11 reference sequence (Derwent AAU82713)

Nucleotide sequence:

```
atgactgtcc gaaacatcgc ctccatctgt aatatgggca ccaatgcctc tgctctggaa     60
aaagacattg gtccagagca gtttccaatc aatgaacact atttcggatt ggtcaatttt    120
ggaaacacat gctactgtaa ctccgtgctt caggcattgt acttctgccg tccattccgg    180
gagaatgtgt tggcatacaa ggcccagcaa aagaagaagg aaaacttgct gacgtgcctg    240
gcggaccttt tccacagcat tgccacacag aagaagaagg ttggcgtcat cccaccaaag    300
aagttcattt caaggctgag aaaagagaat gatctctttg ataactacat gcagcaggat    360
gctcatgaat ttttaaatta tttgctaaac actattgcgg acatccttca ggaggagaag    420
aaacaggaaa aacaaaatgg aaaattaaaa aatggcaaca tgaacgaacc tgcggaaaat    480
aataaaccag aactcacctg ggtccatgag atttttcagg gaacgcttac caatgaaact    540
cgatgcttga actgtgaaac tgttagtagc aaagatgaag attttcttga cctttctgtt    600
gatgtggagc agaatacatc cattacccac tgtctaagag acttcagcaa cacagaaaca    660
ctgtgtagtg aacaaaaata ttattgtgaa acatgctgca gcaaacaaga agcccagaaa    720
aggatgaggg taaaaaagct gcccatggtc ttggccctgc acctaaagcg gttcaagtac    780
atggagcagc tgcgcagata caccaagctg tcttaccgtg tggtcttccc tctggaactc    840
cggctcttca acacctccag tgatgcagtg aacctggacc gcatgtatga cttggttgcg    900
gtggtcgttc actgtggcag tggtcctaat cgtgggcatt atatcactat tgtgaaaagt    960
```

```
cacggcttct ggcttttgtt tgatgatgac attgtagaga aaatagatgc tcaagctatt    1020
gaagaattct atggcctgac gtcagatata tcaaaaaatt cagaatctgg atatatttta    1080
ttctatcagt caagagagta a                                              1101
```

Examples

Specifically known family members of the clan CA/family C12/C19 (deubiquitin enzymes) are retrieved from MEROPS (a total of 9 peptide sequences for C12 and 76 peptide sequences for C19) and they are subject to PSI-BLAST and Smith-Waterman searches. The search databases include Open Reading Frame sequences translated from both Celera and public (NCBI) human genome sequences, Celera predicted proteins, Celera genscan prediction based on Celera human genome, and predicted sequences from proprietary databases. In addition, the members of the two families (C12 and C19) are subjected to Interproscan searches to identify their Interpro motifs. All proteins from the C12 family (Ubiquitin carboxyl-terminal hydrolase, family 1) are identified by IPR001578 and the associated Prints, Prosite and PFAM patterns (PR00707, PS00140 and PF01088). Similarly, family C19 is defined by IPR001394 motif (Ubiquitin thiolesterase) and the corresponding Prints and PFAM signatures (PS00972, PS00973, PS50235, PF00443). Models from the Prosite, Prints and PFAM databases are downloaded and used in the appropriate searches. Results from the various searches were merged and processed by three filters in order to reduce redundancy. Primary hits are screened for low-complexity and are matched against a protein reference database that contains human protein sequences from both GenBank, SwissProt and Refseq. Hits which share a greater than 95% identity over at least 50AA with proteins in the reference database are removed. In addition, the corresponding DNA sequences of the hits are matched against a DNA reference database which contains Refseq human cDNA sequences. Hits which share a greater than 95% identify over at least 150nt with DNA sequences in the reference database are removed. And finally hits that survived the filtering process are purged against each other using a 95% identity cutoff to generate clusters consisting of highly homologous hits. USP_N01 is a representative hit sequence from such cluster.

A second round of analysis is conducted for gene expression profiling and electronic Northern for all human members of the families C12/C19. Public sequences of the C12/C19 family members are used to identify corresponding UniGene cluster and associated normalized expression distribution for tissue types. These are compared with gene chips data obtainable by looking at the corresponding probe sets on a human tissue atlas and on tumor samples .

Example 1:

The gene expression profiles for USP_N01 corresponding to SEQ ID.No. 1 and SEQ ID. No.2 is shown in Figure 1

Example 2:

The gene expression profile for USP_N07 corresponding to SEQ ID No. 5 and SEQ ID. No.6 is shown in Figure 2.

Example 3:

The gene expression profile for USP_N11 corresponding to SEQ ID No. 9 and SEQ ID . No.10 is shown in Figure 3.